DISSERTATIONES SCHOLAE DOCTORALIS AD SANITATEM INVESTIGANDAM
UNIVERSITATIS HELSINKIENSIS

44/2016

MIKKO MUONA

Identification of New Genetic Syndromes with Epilepsy by Whole-Exome Sequencing

INSTITUTE FOR MOLECULAR MEDICINE FINLAND AND
NEUROSCIENCE CENTER AND
RESEARCH PROGRAMS UNIT, MOLECULAR NEUROLOGY
FACULTY OF MEDICINE AND
DOCTORAL PROGRAMME IN BIOMEDICINE
UNIVERSITY OF HELSINKI

FOLKHÄLSAN INSTITUTE OF GENETICS

IDENTIFICATION OF NEW GENETIC SYNDROMES WITH EPILEPSY BY WHOLE-EXOME SEQUENCING

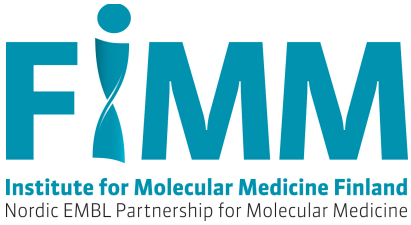
Mikko Muona, MSc

Institute for Molecular Medicine Finland (FIMM),
Research Programs Unit, Molecular Neurology, Faculty of Medicine,
Neuroscience Center,
Doctoral Programme in Biomedicine,
University of Helsinki, Finland
and
Folkhälsan Institute of Genetics, Helsinki, Finland

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine, University of Helsinki, for public examination in Biomedicum Helsinki, Lecture hall 2, Haartmaninkatu 8, Helsinki, on 22 June 2016, at 12 noon.

Helsinki 2016



Dissertationes Scholae Doctoralis Ad Sanitatem Investigandam
Universitatis Helsinkiensis (44/2016)

ISSN 2342-3161 (paperback)
ISBN 978-951-51-2255-1 (paperback)
Hansaprint Oy
Helsinki 2016
ISSN 2342-317X (PDF)
ISBN 978-951-51-2256-8 (PDF)
http://ethesis.helsinki.fi

Supervised by

Professor Aarno Palotie, MD, PhD
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Helsinki, Finland and
The Broad Institute of MIT and Harvard
Boston, MA, USA

Professor Anna-Elina Lehesjoki, MD, PhD
Folkhälsan Institute of Genetics, Helsinki, Finland
and Research Programs Unit, Molecular Neurology
and Neuroscience Center
University of Helsinki, Helsinki, Finland

Thesis advisory committee

Docent Päivi Saavalainen, PhD
Department of Bacteriology and Immunology,
Haartman Institute and
Research Programs Unit, Immunobiology,
University of Helsinki, Helsinki, Finland

Professor Hannes Lohi, PhD
Department of Veterinary Biosciences and Research
Programs Unit, Molecular Neurology
University of Helsinki, Helsinki, Finland and
Folkhälsan Institute of Genetics, Helsinki, Finland

Reviewed by

Docent Marjo Kestilä, PhD
Department of Government Services
National Institute for Health and Welfare
Helsinki, Finland

Helen V. Firth DM, FRCP, DCH
East Anglian Medical Genetics Service,
Addenbrooke's Hospital, Cambridge University
Hospitals
Cambridge, UK and
Wellcome Trust Sanger Institute
Hinxton, Cambridge, UK

Opponent

Professor Joris Veltman
Department of Human Genetics, Donders Centre for
Neuroscience, Radboud University Medical Center,
Nijmegen, the Netherlands and Department of
Clinical Genetics, Maastricht University Medical
Centre, Maastricht, the Netherlands

To my family

ABSTRACT

Epilepsies are a heterogeneous group of central nervous system diseases characterised by recurrent epileptic seizures. They are one of the most common neurological diseases with a lifetime prevalence of ~4%. Epileptic seizures are also a common comorbidity of various neurobiological disorders where epilepsy is not the primary diagnosis. Most epilepsies have a genetic origin, either monogenic or polygenic, however, the causal genetic variants have remained unknown in a substantial proportion of individuals with epilepsies.

Over the past decade, technological advances in DNA sequencing have allowed the characterisation of the genetic basis of human disorders rapidly and efficiently. One of the most widely used methods is whole-exome sequencing (WES) where genetic variants in the protein coding regions of the genome, the exome, are captured. Even though the exome constitutes only ~1.5% of the genome, the majority of disease-causing variants underlying severe, monogenic diseases are located in the protein coding regions.

Here, we aimed to decipher the molecular genetic basis of severe epilepsy syndromes by utilising WES to identify disease-causing genetic variants in patients without a genetic diagnosis. We studied patients with progressive myoclonus epilepsy (PME, n=84) or severe infantile-onset epileptic syndromes (n=30), which are one of the most devastating forms of genetic syndromes with epilepsy and characterised by frequent, pharmacoresistant seizures and poor prognosis. Given that the patients had undergone genetic testing to varying extent prior to this study, we specifically aimed to establish novel genes and molecular biological mechanisms underlying these syndromes.

We made substantial progress in understanding the genetic architecture and molecular basis of the studied syndromes. For PMEs, we established a new major genetic cause and also expanded the genotypic and phenotypic spectrum of previously established disease genes. For severe infantile-onset epileptic syndromes, we identified one new, definite causal gene and one that requires identification of additional patients to confirm the causal role. The three newly identified disease genes represent three different molecular functions that together give new insight on epileptogenic mechanisms.

The new PME subtype is caused by a heterozygous missense variant c.959G>A (p.Arg320His) in *KCNC1* that was identified in 11 unrelated patients (13%) in the PME exome sequencing cohort. We have subsequently identified six additional patients. The gene encodes a potassium ion channel $K_V3.1$ that has an important role in generating action potentials in the central nervous system, with the mutation disrupting the ability to transport potassium ions across the cell membrane. This mutation occurs in most families *de novo,* that is, it is a newly arising mutation. Based on the

estimated mutation rate, the recurrent *KCNC1* mutation is a worldwide cause of PME with likely hundreds of affected individuals globally.

In five families with altogether nine affected siblings, we identified compound heterozygous variants in *UBA5* as the cause of an infantile-onset syndrome characterised initially by irritability, followed by epilepsy, dystonic movements, moderate to severe intellectual disability, microcephaly and stagnation of development. The gene encodes an activating enzyme for UFM1, which is a small ubiquitin-like protein that is conjugated to its target proteins. The function of the highly conserved UFM1 conjugation system is still largely unknown. Functional analysis of the UBA5 mutants suggest that the identified variants cause reduced enzymatic activity of UBA5. Symptoms of the *UBA5* patients and our findings in the central nervous system specific knockout mice for *Ufm1* together indicate that UFM1-cascade is essential for normal development and function of the central nervous system.

Finally, we identified compound heterozygous variants in *ADAM22* as the likely cause of the disease in a patient with an infantile-onset rapidly progressing encephalopathy with epilepsy and cortical atrophy. The gene encodes a postsynaptic protein that functions as a receptor for LGI1, and we show that the identified variants abolish the ability of ADAM22 to bind to LGI1. The LGI1-ADAM22 complex is an antiepileptogenic factor regulating synaptic transmission throughout life. Highlighting the important role of this complex, knockout of *Adam22* and *Lgi1* in mice causes lethal epilepsy. Autosomal dominant *LGI1* variants also cause epilepsy in humans. Identification of a patient with loss-of-function variants in *ADAM22* suggest that also this gene is linked to epilepsy in humans. This connection should be confirmed through identification of additional affected individuals with *ADAM22* variants.

Altogether, this thesis demonstrates the power of WES in identification of causal genetic variants even in phenotypically heterogeneous patient cohorts subjected to prior genetic screenings. The findings improve diagnostics of these syndromes, increase knowledge of the underlying molecular mechanisms and potentially aid in developing new therapeutic interventions. Finally, for these families, establishment of the genetic diagnosis ends years of uncertainty and frustration of not knowing the cause of the disease and prevents need for unnecessary diagnostic testing.

TIIVISTELMÄ

Epilepsiat ovat heterogeeninen joukko keskushermostosairauksia, jotka ilmenevät toistuvina epileptisinä kohtauksina. Niiden elämänaikainen esiintyvyys on noin 4 % eli ne ovat yksiä yleisimmistä neurologisista sairauksista. Epileptisiä kohtauksia esiintyy myös osana muita keskushermoston sairauksia, joissa epilepsia ei ole päädiagnoosi. Useimmat epilepsiat ovat geneettisiä – joko mono- tai polygeenisiä – mutta tautia aiheuttavat geneettiset variantit jäävät tunnistamatta merkittävällä osalla epilepsiaa sairastavista.

Viimeisen vuosikymmenen aikana teknologinen kehitys DNA-sekvensoinnissa on mahdollistanut ihmisen sairauksien geneettisen taustan selvittämisen nopeasti ja tehokkaasti. Yksi käytetyimmistä menetelmistä on eksomisekvensointi, jossa geneettiset variantit koko perimän proteiinia koodaavilla alueilla – eli eksomissa – pystytään tunnistamaan. Vaikka eksomi on vain noin 1,5 % koko perimästä, suurin osa vakavia monogeenisiä sairauksia aiheuttavista muutoksista sijaitsee proteiinia koodaavilla alueilla.

Tässä väitöstutkimuksessa tavoitteenamme oli selvittää vakavien epileptisten oireyhtymien molekyyligeneettistä taustaa hyödyntämällä eksomisekvensointia tautia aiheuttavien varianttien tunnistamisessa potilailla, joilla ei ole geneettistä diagnoosia. Tutkimusaineistomme koostui potilaista, joilla on joko progressiivinen myoklonusepilepsia (PME, n=84) tai vakava imeväisikäisenä alkava epileptinen oireyhtymä (n=30). Nämä oireyhtymät kuuluvat vakavimpien epileptisten oireyhtymien joukkoon ja niihin liittyy toistuvia, lääkeresistenttejä kohtauksia ja niillä on huono ennuste. Koska aineiston potilaille oli tehty geneettisiä diagnostisia testejä ennen tutkimukseen osallistumista, ensisijaisena tavoitteenamme oli tunnistaa uusia tautigeenejä ja molekyylimekanismeja näiden oireyhtymien taustalla.

Edistimme tutkimuksen avulla merkittävästi ymmärrystämme näiden oireyhtymien molekyyligeneettisestä taustasta. Tunnistimme uuden PME-alatyypin ja kaksi uutta autosomaalisesti peittyvästi periytyvää imeväisikäisten vakavaa enkefalopatiaa, joista toisen kohdalla vaaditaan vielä lisäpotilaiden tunnistamista varmistuaksemme havaitsemiemme varianttien patogeenisyydestä. Lisäksi laajensimme aiemmin tunnistettujen tautigeenien genotyyppi- ja fenotyyppikirjoa. Nämä kolme uutta tautigeeniä edustavat kolmea eri molekulaarista mekanismia, jotka yhdessä lisäävät tietoa epilepsioiden taustalla olevista tekijöistä.

Väitöskirjan ensimmäisessä osatyössä tunnistamamme uusi PME-alatyyppi aiheutuu heterotsygoottisesta missense-variantista c.959G>A (p.Arg320His) geenissä *KCNC1*. Tämä muutos on 11 potilaalla (13 %; eivät toisilleen sukua) PME-eksomisekvensointiaineistossa. Alkuperäisen löydöksen jälkeen olemme tunnistaneet kuusi lisäpotilasta. *KCNC1* koodaa

kaliumionikanavaa $K_V3.1$, jolla on keskushermostossa tärkeä tehtävä aktiopotentiaalien muodostamisessa. Tunnistamamme mutaatio tässä geenissä vahingoittaa kanavan kykyä siirtää kaliumioneita solukalvon läpi. Mutaatio on niin kutsuttu *de novo* –mutaatio eli se on uusi muutos, joka ei ole periytynyt vanhemmilta. Arvioimamme mutaatiotaajuuden perusteella tämä mutaatio *KCNC1*-geenissä aiheuttaa PME:n sadoilla potilailla maailmanlaajuisesti.

Toisessa osatyössä tunnistimme viidessä perheessä yhteensä yhdeksällä potilaalla yhdistelmäheterotsygootit *UBA5*-geenin variantit, jotka aiheuttavat imeväisikäisenä ilmenevän oireyhtymän. Tämä oireyhtymä ilmenee aluksi ärtyvyytenä, ja muita myöhemmin esiintyviä oireita ovat epileptiset kohtaukset, dystoniset liikkeet, älyllinen kehitysvamma ja pienipäisyys. UBA5 on UFM1-proteiinia aktivoiva entsyymi. UFM1 on pieni ubikitiinin kaltainen proteiini, joka kiinnitetään sen kohdeproteiineihin entsyymien katalysoimien reaktioiden kautta. UFM1-kaskadin tehtävä soluissa on suurelta osin tuntematon. Funktionaaliset kokeemme viittaavat siihen, että tunnistamamme *UBA5*-variantit vähentävät UBA5:n entsymaattista aktiivisuutta. *UBA5*-potilaiden oireet ja löydökset tutkimallamme *Ufm1*-hiirimallilla osoittavat, että UFM1-kaskadi on välttämätön keskushermoston kehittymiselle ja toiminnalle.

Kolmannessa osatyössä tunnistimme yhdistelmäheterotsygoottiset variantit *ADAM22*-geenissä todennäköisenä aiheuttajana potilaan imeväisikäisenä alkaneelle etenevälle enkefalopatialle, johon liittyy epileptisiä kohtauksia ja aivokuoren atrofiaa. ADAM22-proteiini toimii LGI1:n reseptorina ja osoitimme tutkimuksessa, että tunnistamamme variantit estävät ADAM22-proteiinin ja LGI1:n välisen interaktion. LGI1-ADAM22 –kompleksi on antiepileptinen tekijä, joka säätelee synaptista transmissiota läpi elämän. Näiden proteiinien tärkeää tehtävää korostaa se, että *Lgi1*- ja *Adam22*-poistogeeniset hiiret kärsivät epileptisistä kohtauksista ja kuolevat 2-3 viikkoa syntymänsä jälkeen. Autosomaalisesti vallitsevasti periytyvät variantit *LGI1*-geenissä aiheuttavat epilepsiaa ihmisellä. Vakavasta epileptisestä oireyhtymästä kärsivällä potilaalla tunnistamamme proteiinin toiminnan estävät, peittyvästi periytyvät variantit *ADAM22*-geenissä viittaavat siihen, että myös *ADAM22* kytkeytyy epilepsiaan ihmisellä. Tämän kytköksen vahvistaminen edellyttää lisäpotilaiden tunnistamista.

Kokonaisuudessaan tämä väitöstutkimus osoittaa eksomisekvensoinnin tehokkuuden tautia aiheuttavien varianttien tunnistamisessa heterogeenisissä potilaskohorteissa, joille on tehty aiempia diagnostisia testejä. Löydöksemme edistävät näiden epileptisten oireyhtymien diagnostiikkaa, lisäävät tietoa niiden taustalla olevista molekyylimekanismeista ja mahdollisesti auttavat kehittämään uusia hoitokeinoja. Perheille diagnoosin saaminen merkitsee myös sairauden aiheuttajaan liittyneen epätietoisuuden ja turhautuneisuuden päättymistä ja tarpeettomien diagnostisten testien loppumista.

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications:

I **Muona M**, Berkovic SF, Dibbens LM, Oliver KL, Maljevic S, Bayly MA, Joensuu T, Canafoglia L, Franceschetti S, Michelucci R, Markkinen S, Heron SE, Hildebrand MS, Andermann E, Andermann F, Gambardella A, Tinuper P, Licchetta L, Scheffer IE, Criscuolo C, Filla A, Ferlazzo E, Ahmad J, Ahmad A, Baykan B, Said E, Topcu M, Riguzzi P, King MD, Ozkara C, Andrade DM, Engelsen BA, Crespel A, Lindenau M, Lohmann E, Saletti V, Massano J, Privitera M, Espay AJ, Kauffmann B, Duchowny M, Moller RS, Straussberg R, Afawi Z, Ben-Zeev B, Samocha KE, Daly MJ, Petrou S, Lerche H, Palotie A, Lehesjoki A-E (2015). A recurrent *de novo* mutation in *KCNC1* causes progressive myoclonus epilepsy. *Nature Genetics* 47:39-46.

II **Muona M***, Ishimura R*, Laari A*, Ichimura Y, Linnankivi T, Keski-Filppula R, Herva R, Rantala H, Paetau A, Pöyhönen M, Obata M, Uemura T, Karhu T, Bizen N, Takebayashi H, McKee S, Parker MJ, Akawi N, McRae J, Hurles ME, the DDD study, Kuismin O, Kurki MI, Anttonen A-K, Tanaka K, Palotie A, Waguri S, Lehesjoki A-E*, Komatsu M*. Biallelic variants in *UBA5* link dysfunctional UFM1 ubiquitin-like modifier pathway to severe infantile-onset encephalopathy. *In revision*.

*These authors contributed equally.

III **Muona M**, Fukata Y, Anttonen A-K, Laari A, Palotie A, Pihko H, Lönnqvist T, Valanne L, Somer M, Fukata M, Lehesjoki A-E (2016). Dysfunctional ADAM22 implicated in progressive encephalopathy with cortical atrophy and epilepsy. *Neurology: Genetics* 2**:**e46.

The publications are referred to in the text by their roman numerals. The original publications are reproduced with the permission of their copyright holders.

AUTHOR'S CONTRIBUTIONS

I-III: Participated in the study design. Established analysis pipeline for the whole-exome sequencing data. Processed the sequence data and generated variant calls and annotations. Analysed the variant data. Interpreted the functional and clinical significance of the identified variants, with contributions from co-authors. Performed a subset of variant validations and segregation analyses and analysed the data. Established hypotheses to be tested in functional assays for identified variants and interpreted results of functional experiments together with co-authors. In study II, performed molecular assays to examine the effect of one of the identified variants on gene splicing. Drafted the manuscripts and finalised them after receiving comments from co-authors.

ABBREVIATIONS

A	adenosine	ILAE	International League Against Epilepsy
ACMG	American College of Medical Genetics and Genomics	indel	insertion or deletion variant
bp	base pair	kb	kilobases
C	cytosine	Kv	potassium voltage-gated channel
CCDS	Consensus coding sequence	MEAK	myoclonus epilepsy and ataxia due to potassium channel mutation
cDNA	complementary DNA	MRI	magnetic resonance imaging
CNS	central nervous system	mtDNA	mitochondrial DNA
CNV	copy-number variant	NCBI	National Center for Biotechnology Information
DDD	Deciphering Developmental Disorders study	NCL	neuronal ceroid lipofuscinosis
EE	epileptic encephalopathy	NGS	next-generation sequencing
EEG	electroencephalogram	RT-PCR	reverse transcription polymerase chain reaction
ER	endoplasmic reticulum	PME	progressive myoclonus epilepsy
EVS	Exome Variant Server	SISu	Sequencing Initiative Suomi
ExAC	Exome Aggregation Consortium	SNV	single-nucleotide variants
FIMM	Institute for Molecular Medicine Finland	T	thymine
G	guanine	UBL	ubiquitin-like protein
GABA	gamma-aminobutyric acid	ULD	Unverricht-Lundborg disease
GATK	Genome Analysis Toolkit	WES	whole-exome sequencing
GGE	genetic generalised epilepsy	WGS	whole-genome sequencing
GWAS	genome-wide association study	WTSI	Wellcome Trust Sanger Institute

1 INTRODUCTION

The pace of technological advances in the field of genetics during the past decade is unrivalled in almost any other sector in science. The new sequencing technologies allow the characterisation of the human genome and its variation in the highest detail and in large numbers of individuals with relatively low price. Given that the majority of human disorders bear a genetic component, it has generated a lot of hope and hype that increased understanding of the disease-associated genetic variation would translate into clinical applications where diagnosis and treatments would be more efficient and accurate. Personalised medicine is the buzzword that underlies the overall aim of the current medical research to be able to identify the causal molecular mechanisms in each individual and offer tailored treatments based on the findings.

Rare diseases have been on the front line of reaching individual-level molecular diagnoses. This is largely due to their less complex genetic architecture with genetic defects occurring mostly in single genes only. While many genes underlying rare diseases were identified using traditional approaches, latest sequencing technologies have made it possible to solve cases that were a challenge to earlier methods and thus boosted the gene discovery substantially (Boycott *et al.* 2013). However, there is still much work to do to achieve complete understanding of the genetic basis of rare diseases; the underlying genetic defects remain to be characterised in approximately half of the 7,000 rare diseases with genetic origin (Chong *et al.* 2015). The goal to provide targeted therapies in these diseases is even further, since only a minority of the rare diseases have a specific drug used in the treatment. It is important to achieve these goals in rare diseases; they are often severe and life threatening and even though individually rare, they collectively affect more than 5% of the population and thus represent a major socio-economic burden (Chong *et al.* 2015).

Epilepsies, which are a group of diseases characterised by recurrent epileptic seizures due to aberrant brain activity, represent one of the clinical entities of which aetiology has remained largely uncharacterised before the recent advances in the genetic and neurobiological research (Noebels 2015; Staley 2015). Considering the fact that epilepsies were long regarded as magical or divine diseases, it is fascinating that genetic studies of epilepsies are now actually one of the greatest success stories in terms of the number of new gene discoveries (Myers and Mefford 2015). While majority of the epilepsies remain genetically unsolved, identification of the underlying genetic defects in many forms of epilepsies has finally provided comprehensive insight on what molecular mechanisms trigger seizure activity in the brain.

In this study, we utilised modern sequencing technologies to dissect the genetic basis of unexplained, rare syndromes with epilepsy. Our focus was on

the most devastating forms of epileptic syndromes that occur in early life, do not respond to treatments and where symptoms get worse over time. Identification of the genetic cause of the disease in these patients would end the ‘diagnostic odyssey’, which typically involves years of frustration, numbers of medical appointments and unsuccessful diagnostic tests. Moreover, establishing the genetic diagnosis is the first step on the way to find targeted therapies to the disease.

2 REVIEW OF THE LITERATURE

2.1 Variation in the human genome

The human genome comprises of the nuclear and the mitochondrial genome. The former consists of over 3 billion base pairs (bp), either adenosine (A) pairing with thymine (T) or cytosine (C) with guanine (G). In the nuclear genome, the letters of DNA are subdivided into 22 pairs of autosomes and one pair of sex chromosomes. The mitochondrial genome is only 16.6 kilobases (kb) in size but is present in multiple copies in each mitochondrion. Genes are functional elements of the genome. Recent estimates suggest that the human genome contains approximately 19,000-20,500 protein-coding genes (Clamp *et al.* 2007; Ezkurdia *et al.* 2014). However, splicing of messenger RNA, which shuffles the coding parts of the genes before translation into amino acid sequence, and post-translational modifications of proteins, increase the true number of unique protein molecules present in cells to a many folds higher level. In addition, the human genome contains RNA genes, which possess information of a wide array of functional RNA elements, many of which have only recently been discovered.

Together with environmental factors, genetic information embedded in the human genome determines the characteristics of each individual. Variation in the DNA sequence underlies traits, such as height and eye colour, but also disorders. A major step towards a better understanding of the role genetics in human health and disease was completion of the sequence of the human genome over a decade ago (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001; International Human Genome Sequencing Consortium 2004). The main product of these efforts was the reference genome, to which all newly generated sequence data is compared. After solving the sequence of the human genome, a number of large-scale genotyping and sequencing efforts have generated publicly available catalogues of all types of genetic variation across various human populations (e.g., International HapMap Consortium 2003; The 1000 Genomes Consortium 2010). This information has provided us basic information of the nature of the genetic variation within and between populations. Importantly, the generated data are invaluable in analysis of the genetic basis of both common and rare diseases, since they help us to understand which fraction of the variation is common and likely benign and which variation may be associated with disease.

Mutations are permanent changes in the DNA sequence of the genome and thus are the source of genetic variation. Differences between the nucleotide sequence of any newly generated genetic data and that in the reference genome can be reported as genetic variants. Genetic variation ranges from single nucleotide level changes to gains and losses of whole chromosomes. Single-nucleotide variants (SNVs) are changes of one DNA

letter to another. Small insertions and deletions (indels) are small (<50bp) gains or loss of genetic material. Copy-number variants (CNVs) in turn are larger scale duplications or deletions of sections of DNA. Other forms of structural variants include inversions and translocations.

Numerically SNVs and indels are most abundant, but CNVs affect a bigger number of base pairs (The 1000 Genomes Consortium 2010; Campbell and Eichler 2013). Small-scale changes in the genome can arise due to spontaneous mutations (e.g. errors in DNA replication) or induced mutations (exposure to external mutagens such as ultraviolet radiation). Normally these errors are repaired by dedicated molecular pathways, but failures in these processes lead to permanent sequence variants that can be transmitted to next cell generations, if they occur in somatic cells, or to offspring, if they take place in germline cells. CNVs are mostly caused by recombination between mispaired sequences during meiosis (Campbell and Eichler 2013). Mutations do not occur evenly across the genome. There are so called mutation hotspots in which mutations take place more often. For example, CG-dinucleotides (CpG sites) are more prone to C to T transitions than other nucleotide pairs: the estimated rate of germline base substitutions is 1.2×10^{-8} per base per generation, equalling to an average of 63 new mutations per person, but in CpG sites, the rate is 1.12×10^{-7} (Kong *et al.* 2012). Importantly, germline mutations are mostly of paternal origin (Kong *et al.* 2012).

2.2 Genetic variation in human disease

Each human genome contains ~4 million SNVs and indels, of which ~25,000 occur in protein coding regions (The 1000 Genomes Consortium 2012). Altogether, over 100 million variant sites have been catalogued in dbSNP, a database for human sequence variation. Already several decades ago it was hypothesized that most of the genetic variation within and across species are neutral and does not have a dramatic evolutionary effect (Kimura 1968). Along this line, recent data from genome sequencing projects have shown that the human genome can tolerate a substantial degree of variation without any apparent pathogenic effect. For example, a typical human genome contains over 100 rare loss-of-function variants that are predicted to disrupt the gene function and a subset of them even affect both copies of the gene (The 1000 Genomes Consortium 2010; Sulem *et al.* 2015). Highlighting the widespread and dense nature of genomic variation, sequencing of the protein-coding elements of the genome (exome) in 60,000 individuals as part of Exome Aggregation Consortium (ExAC), showed that there is a variant in every 8 bp in at least one individual of the sample set (Lek *et al.* 2015).

However, not every kind of genetic variation is neutral or without phenotypic effect. Some variants increase susceptibility to certain disorders or even directly cause them. Moreover, some variants are evolutionary

advantageous and are subjected to positive selection. In gene level, genes that have, for example, highly important function in the embryonic development tolerate only a little variation in the amino acid sequence, whereas some other genes, such as those encoding olfactory receptors are more likely to be nonessential and can more often be completely knocked out without obvious effect (Sulem *et al.* 2015). Given that proteins represent important functional elements of the cell, it is therefore no surprise that much of the deleterious variation that is eliminated by negative, purifying selection resides within the exome. This can be seen in the sequence data from the 1000 genomes project, which showed that gene exons have 50% less genetic diversity compared to intronic regions (The 1000 Genomes Consortium 2010). Particularly during the recent years, however, the importance of variation in noncoding regions of the genome for human disease has also been noted. For example, many variants that predispose to multifactorial diseases are located in regulatory regions of genes and affect the levels of gene expression (Albert and Kruglyak 2015). Noncoding variation can also have a dramatic effect on gene function, for example by disrupting splicing (Epstein 2009).

Genetic variants can be classified based on their consequence on the function of the gene product. Loss-of-function variants result in abolished function. Hypomorphic variants are a subtype of this class and cause a partial loss of function. Dominant-negative variants have the ability to disrupt the function of the wild-type allele. Gain-of-function variants result in an enhanced or altered function, for example, an ion channel remains always open. More specifically, variants can be classified based on their effect on protein sequence. Silent or synonymous variants do not change the amino acid sequence, whereas missense variants cause a substitution of an amino acid residue to another. Variants that are commonly categorised under the name of loss-of-function variants include splice site (disrupt splicing), nonsense (premature stop codon) and frameshift indel (disrupt reading frame) variants. However, these variants do not necessarily always cause loss of function of a gene, for example, when a nonsense mutation occurs in the very end of the coding sequence of the gene. On the other hand, one needs to bear in mind that also noncoding, missense or synonymous variants can have a complete loss-of-function effect if, for example, a missense change disrupts a catalytically active residue or a synonymous change disrupts splicing.

2.3 Genetics of rare diseases

By European Union definition, rare diseases have a prevalence less than 1 in 2000 in the population (European Comission 2016). It is difficult to estimate precisely the number of rare diseases but a commonly cited figure is around 6,000-7,000 (Boycott *et al.* 2013; Chong *et al.* 2015). In aggregate, rare diseases are not uncommon. As many as 25 million of the US population are affected by a rare disease and globally this number reaches 350 million

(Chong *et al.* 2015; Global Genes 2016). Rare diseases are often severe, chronic and without an effective treatment. Moreover, their diagnosis is challenging due to a number of factors, such as their rarity (clinicians may not have ever seen a patient with that condition) and clinical overlap with other more common ailments. Consequently, a substantial proportion of affected individuals receive an incorrect diagnosis initially and getting the correct diagnosis may take several years (average 4.8 years) or in the worst case it is never reached (Shire 2016). Because of the strong genetic component underlying most rare diseases, it is important to utilise genetic technologies to improve the currently suboptimal diagnosing of rare diseases. Getting an accurate diagnosis quickly is important for a number of reasons. First, it reduces emotional toll of the patient and family of not knowing what causes the severe illness and assists family counselling when mode of inheritance is known. Knowing the specific cause may assist selecting optimal treatments and give insight on prognosis. Moreover, it eliminates need for unnecessary diagnostic tests involving numerous medical appointments and high costs. Finally, with the diagnosis the family can be in contact with others who are dealing with the same disease and seek peer support.

Rare genetic diseases are mostly Mendelian, which means that their segregation follows the simple rules of inheritance proposed by Gregor Mendel in the 19th century. This is because Mendelian (or monogenic) diseases are caused by genetic defects that are typically fully penetrant and reside in single genes. In autosomal recessively inherited Mendelian diseases, disease develops if deleterious variants affect both copies of the gene. In the case of X-linked recessive disorders, males develop a disease if their only copy of the X chromosome harbours the variant. Recessive diseases are typically caused by loss-of-function or hypomorphic variants. Dominantly inherited Mendelian diseases are caused by heterozygous variants that can be gain-of-function, loss-of-function or dominant-negative. Dominant disease genes where even a heterozygous loss-of-function variant is enough to cause a disease are called haploinsufficient because expression from both copies of the gene is required for normal cell function. Recent large-scale sequencing efforts have been able to pinpoint genes that do not tolerate deleterious variation in even one copy of the gene. By looking at the amount of deleterious variation (missense, loss-of-function) in exome sequences of the general population, researchers have recognised genes that have less variation than known estimates on mutation rates would suggest, indicating that heterozygous variants in these genes are under stronger negative selection (Petrovski *et al.* 2013; Lek *et al.* 2015). Indeed, many previously known disease genes where heterozygous variants cause a disease rank among the least tolerant genes for functional variation.

In addition to inherited genetic factors, *de novo* mutations, i.e. newly arising mutations occurring either in parental germline cells or at some point after conception, contribute to human diseases. Many rare diseases are so

severe that patients cannot reproduce and causal variants are then not transmitted to next generations. In these disorders, spontaneously occurring *de novo* mutations and/or recessively inherited variants mostly explain their presence in the human population (Veltman and Brunner 2012). Besides rare, severe diseases (Vissers *et al.* 2010), the importance of *de novo* variation in more common and genetically more complex conditions such as schizophrenia or autism, has been recently understood (Xu *et al.* 2011; O'Roak *et al.* 2012; Campbell and Eichler 2013). These observations suggest that at least some forms of these complex disorders may be caused, or at least contributed, by *de novo* genetic defects in single genes rather than by common variation in multiple genes.

Somatic mutations are *de novo* genetic alterations that occur after conception in any cell type but germ cells. Consequently, they do not affect all but a subset of cells in the body and are not inherited. Their role in human disease beyond cancer has only recently started to be acknowledged. For example, a variant may be present in a large proportion of neurons, but it still missed in genetic testing because it is not present at all or at lower levels in peripheral blood, which is typically used as the source of DNA in genetic studies. Interestingly, recent studies with access to postmortem samples or brain biopsies obtained by surgery have been able to characterise somatic variation in the brain and shown that somatic mutations may cause brain malformations and epilepsy (Poduri *et al.* 2013; Lodato *et al.* 2015).

The developmental stage when a newly arising mutation occurs has important implications regarding family counselling and estimating the recurrence risk of diseases caused by *de novo* alterations (Vadlamudi *et al.* 2010; Veltman and Brunner 2012). If the mutation occurs in a single gamete or after conception, the recurrence risk of the disease in the family is close to 0%. However, if the mutation affects some if not all diploid germline cells of one of the parents the risk can be up to 50%. At individual level a specific recurrence risk is not typically possible to obtain, but deep sequencing of parental blood samples for the mutations have shown promise in providing more accurate estimates in individual families (Rahbari *et al.* 2016). On average, the recurrence risk of diseases caused by *de novo* mutations has been recently estimated to be approximately 1.3% (Rahbari *et al.* 2016).

2.3.1 Traditional and modern techniques of genetics research in Mendelian disease

In addition to the benefits related to the diagnosis and care of the patients, a better understanding of the genetic basis of Mendelian disease genes has also more general outcomes. Identification of a new disease gene is often associated with increased knowledge of the gene function and related cellular processes (Boycott *et al.* 2013). Insight on rare diseases can also enhance our understanding of related common, complex disorders (Peltonen *et al.* 2006). It has been estimated that the underlying genes have not yet been identified in 50% of all known Mendelian diseases (Boycott *et al.* 2013; Chong *et al.*

2015) and in addition many new diseases remain to be discovered. It has been suggested that the human genome contains up to 7,000-15,000 disease genes (Cooper *et al.* 2010a), so likely many thousands of new gene-phenotype associations are to be made. This hypothesis is supported by findings in the ExAC data where over 3,000 genes are almost completely depleted of truncating variants and most of these genes have not been associated with human disease (Lek *et al.* 2015).

Given their more simple genetic background compared to multifactorial, complex conditions, Mendelian diseases have been on the front line of gene discovery for long time. Disease gene mapping, i.e., localisation of genes with disease-causing variants, initiated in 1980s with the discovery of the gene underlying Huntington's disease (Gusella *et al.* 1983). For the next two decades, gene discovery studies relied primarily on genome-wide linkage analysis where disease genes were mapped using polymorphic genetic markers that are in linkage with the disease gene locus, i.e. the marker and the gene locus segregate together. Other methods used successfully include karyotyping and homozygosity mapping (Gilissen *et al.* 2011). While these methods allowed identification of numerous disease loci, the genetic findings were mainly limited to familial diseases that were clinically distinguishable. Moreover, disease gene mapping studies in isolated populations, such as Finland, were largely successful, since these studies benefited from the reduced genetic heterogeneity resulting from genetic bottlenecks in the early phases of population history (Peltonen *et al.* 1999). Traditional genetic methods were poor in dissecting the genetic background of complex diseases because the effect sizes of the underlying genetic factors are small compared to Mendelian diseases and also because the traditional methods do not tolerate high levels of genetic or phenotypic heterogeneity (Baron 2001).

The gold-standard method for DNA sequencing was for many years the dideoxy chain-termination method initially developed by Frederick Sanger in the 1970s (Sanger *et al.* 1977). For example, the first human genome was sequenced using this technology that is usually called 'Sanger sequencing'. However, the capacity of Sanger sequencing is not sufficient for high-throughput genome-level sequencing, illustrated by the $3 billion budget of sequencing the first human genome (Wetterstrand 2016). Still, Sanger sequencing is widely used but its applications are limited to small-scale tasks such as confirmation of variants identified using other methods.

In 2005, the first commercial sequencing machine of the 'next-generation sequencing' (NGS) era was introduced. NGS machines are able to generate low-cost, high-throughput sequence data. Their basic principle is to sequence clonally amplified or single molecule templates in massively parallel fashion (Rehm *et al.* 2013). There are a number of chemistries underlying various NGS technologies but one of the most widely used is the sequencing-by-synthesis method originally developed at Solexa, which was later acquired by Illumina. This method utilizes fluorescently labelled nucleotides that are added one-by-one by a DNA polymerase to amplify single-stranded DNA

molecules attached to the surface of a flow cell (Mardis 2013). An image is taken of each labelled nucleotide that is incorporated to the DNA molecule. Based on the image, a computer determines the correct base at each position, and the base calls are used to deduce the nucleotide sequence. This method produces sequence reads that are relatively short, 50-300 bp in length, which makes it more challenging to align the sequence reads to the reference genome and limits its usability to detect for example long repeat polymorphisms and complex genomic rearrangements. However, the issues caused by short read length can be reduced by using paired-end sequencing where each DNA molecule is sequenced from both directions.

Illustrating the rapid progress of NGS technologies, we have reached a stage where the actual sequencing is not anymore the rate-limiting step in genomics research. The production cost of sequencing one human genome can now be as low as $1,000 in large-scale sequencing facilities (downstream informatics costs not included in the figure), and instead of years that it took to sequence the first human genome, it takes now less than a week (Wetterstrand 2016). Now, the more costly and challenging part is the storage, processing and interpretation of the massive amount of genetic information that is now being generated at the rate of thousands and thousands of whole-exomes and genomes per year.

2.3.2 Gene discovery and diagnostics in the next generation sequencing era

The introduction of NGS technologies has had important implications on the gene discovery in human diseases. Instead of using genetic markers as proxies to the disease-causing variants, genetic variation can be now detected directly for many individuals in parallel. Indeed, the modern approaches have boosted gene discovery in Mendelian diseases (Boycott *et al.* 2013). It is now possible to decipher the genetic basis of diseases showing clinical and genetic heterogeneity and that of diseases that are mostly sporadic, which was challenging with traditional methods.

One of the main advantages of NGS-based methods in genetic studies is that they enable utilisation of 'genotype-first' or 'reverse phenotyping' approach (**Figure 1**). It means that the usual diagnostic approach relying primarily on the clinical features of the patient can be reversed and the precise diagnosis can be established based on the genetic findings (Stessman *et al.* 2014; de Goede *et al.* 2016). This is based on the ability of NGS technologies to examine genetic variation in a large number of individuals in an exome- or genome-wide manner without need for *a priori* hypotheses regarding the biological function of disease genes. The following example illustrates the benefits of the genotype-first approach: let us imagine a study cohort of 100 cases, who belong to a genetically and phenotypically heterogeneous disease entity. With NGS, one could recognise that 10 cases of the cohort have pathogenic variants in the same gene, so they have the same molecular diagnosis. However, if diagnosis had to rely solely on clinical

symptoms, in the absence of specific symptoms, it could have been virtually impossible to recognise the etiological subgroups. Furthermore, with traditional genetic methods it would have been difficult to identify the 10 cases with genetically identical disease because of the genetic heterogeneity in the cohort. One of the advantages of the genotype-first approach is also that it can be used to make a specific diagnosis in cases where extensive clinical phenotyping is not possible (Shi *et al.* 2014). Furthermore, hypothesis-free, genome-wide approaches allow expansion of the phenotypic spectrum of already known disease genes. For example, *TBC1D24* is a gene that was originally linked to epilepsy, but it has since reported also in deafness and a multi-organ syndrome (Falace *et al.* 2010; Campeau *et al.* 2014; Rehman *et al.* 2014).

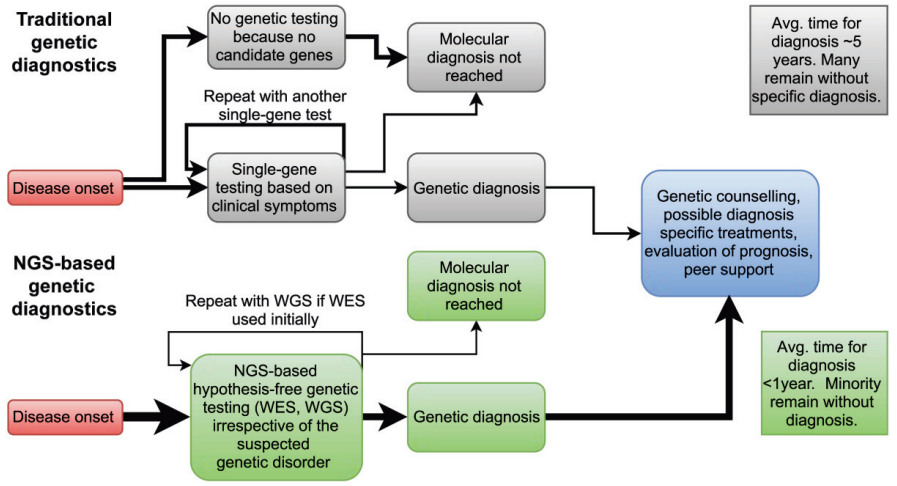


Figure 1 Genetic diagnostics before and during the next-generation sequencing era. The figure illustrates how next-generation sequencing (NGS) methods, whole-exome sequencing (WES) and whole-genome sequencing in particular (WGS; see below section 2.3.3), have the ability to fasten the diagnostic process of rare diseases and increase the proportion of genetically diagnosed patients. Line and arrow width indicates the proportion of patients falling into that category.

2.3.3 Next-generation sequencing applications in gene discovery and diagnostics of rare diseases

There are three main NGS-based applications in human genetics: gene panels, whole-exome sequencing (WES) and whole-genome sequencing (WGS) (**Table 1**).

In gene-panel sequencing, a selected set of genes, ranging from one to hundreds of genes, are sequenced. Selected genes are usually those where pathogenic variants have been previously reported for the disease entity in question. Typically, the coding regions of the genes are targeted. This approach suits well to diagnostic applications when the clinical phenotype suggests a specific genetic disorder with known disease gene(s) so it is more

cost-efficient to do targeted sequencing in few genes. Moreover, gene panels can be a cost-efficient method to screen a replication cohort for variants in candidate genes identified in a discovery sample set (e.g., D'Alessandro *et al.* 2015). It is also a practical solution when extremely high sequencing coverage is needed to target certain clinically relevant genes, for instance, in tumour sequencing where somatic mutations may occur at low levels (Tripathy *et al.* 2014). The negative side of the gene panel approach is obviously that it cannot be used to identify new disease genes, unless one selects candidate genes to the panel based, for example, on function (Carvill *et al.* 2013a; Syrbe *et al.* 2015).

Table 1. Comparison of next-generation sequencing applications.

Property	Gene panels	WES	WGS
Number of genes covered	1-500	20,000	20,000
Bases covered	Variable	~50 Mb	~3Gb
Typical mean sequencing coverage	>100×	60-120×	30-60×
Completeness of target region coverage	++(+)	+(+)	++(+)
Screening of known genetic causes	+++	++	++
Identification of new genes	-	++	+++
Detection of noncoding variants	-	-	+++
Detection of CNVs and other structural variants	-	+	+++
Cost	++	+	-

WES, whole-exome sequencing; WGS, whole-genome sequencing; Mb, megabases; Gb, gigabases; CNV, copy number variant.
Plus and minus symbols are used to score the utility of the different NGS methods in various categories.

When one aims to discover new gene-disease associations, a more extensive approach targeting all genes is required (Rehm *et al.* 2013). Exome- or genome-wide sequencing may also be selected if 1) gene-panel testing is negative, 2) the phenotype of the patient is unspecific and selecting a correct gene panel is demanding, or 3) if there are hundreds of genes underlying the disorder in question, such as intellectual disability.

As mentioned earlier, majority of variants causing Mendelian diseases reside within the exome (Chong *et al.* 2015). Therefore, during the past few years the most common approach to identify novel disease-causing variants has been to perform WES where the sequencing library is created using RNA or DNA bait probes to designed capture exonic and splice site regions (**Figure 2**). Since the exome constitutes of only about 1.5% of the whole genome, WES is a cost-efficient method ($500-1500) to tackle the most relevant parts of the genome in terms of potentially deleterious variation. The first application of WES in gene discovery was published in 2009 (Ng *et al.* 2010). Thereafter, WES has proven to be successful to dissect the genetic basis of Mendelian disorders of all possible mode of inheritance (reviewed by

Boycott *et al.* 2013). The typical diagnostic yield of exome sequencing in large-scale studies of Mendelian disorders is typically around 25-30%, which is higher compared to conventional genomic assays such as karyotyping and CNV analysis with microarrays (reviewed by Chong *et al.* 2015). However, the true diagnostic yield of WES is likely higher because many of the study subjects of the current WES studies have already underwent other genetic testing prior to the study. Indeed, in a recent study where WES was used as a first-tier genetic test in infants with suspected Mendelian diseases, diagnostic yield was 58%, whereas application of single- or multigene panels to the same individuals led to genetic diagnosis in only 14% of cases (Stark *et al.* 2016). WES has also contributed to the understanding of cancer and other complex conditions (Rabbani *et al.* 2014; Van Allen *et al.* 2014). However, because individual genetic factors have only a minor effect on susceptibility of common disorders, large sample sets of thousands of individuals are required in order to identify statistically significant genetic associations in complex conditions (Do *et al.* 2012). This requirement has hindered utility of WES and other NGS applications in common, complex diseases. In general, one of the limitations of WES is that it does not typically cover all target regions with adequate sequencing depth (5-10%), and consequently clinically important variants may be missed. GC-rich regions are in particular challenging regions to capture (Asan *et al.* 2011). However, with most recent versions of exome capture kits, this issue has become less substantial (Lelieveld *et al.* 2015). While WES can be used to detect CNVs (Fromer *et al.* 2012; Poultney *et al.* 2013), the sensitivity and resolution is lower than in other approaches.

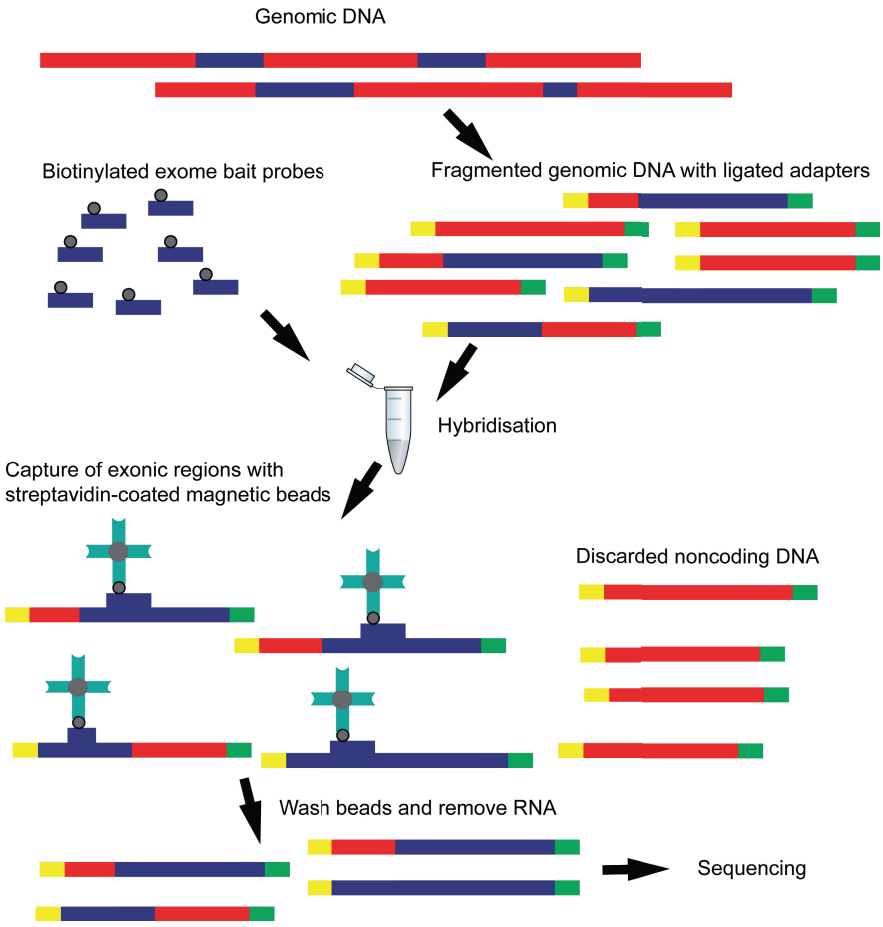


Figure 2 Whole-exome target capture protocol. This protocol applies to exome capture methods based on in-solution hybridisation, which is the most widely used approach. First, genomic DNA is fragmented, and adapters (yellow and green) are ligated to DNA fragments (blue colour indicates exonic regions, red noncoding). Next, biotinylated RNA baits designed to bind to exonic regions are hybridised with the DNA library. Streptavidin-coated magnetic beads attach to biotin on the probes and the complexes are pulled down using magnets. Finally, beads are washed away, RNA digested and exonic DNA subjected to sequencing.

With WGS, it is possible to detect virtually all SNVs, indels and CNVs in one assay, which separates it from other genetic tools. WGS may even cover variants in exonic regions better than WES because it provides more uniform sequencing coverage and more balanced allele ratios (Meynert *et al.* 2014; Lelieveld *et al.* 2015). The caveat of WGS, and other NGS technologies that rely on short-read sequencing technologies, is that they cannot detect repeat expansions, which cause, for example, many neurodegenerative diseases (Everett and Wood 2004). Its usage has also been prohibited by high cost (2-4 times more than WES without taking additional computational burden into account (Lelieveld *et al.* 2015)), but as described above, with the rapid

decline in costs it is becoming now a more feasible application in Mendelian genetics (Gilissen *et al.* 2014; Taylor *et al.* 2015; Willig *et al.* 2015; Stavropoulos *et al.* 2016). The utility of WGS is also affected by the limited knowledge about the functional consequence of variants outside the coding region. Tools aiming to surmount this issue have, however, been developed recently (Kircher *et al.* 2014; Ionita-Laza *et al.* 2016).

Altogether, recent studies have demonstrated that WES and more recently WGS provide a powerful method to diagnose genetic disorders. The cost of a single run of WES, and in particular WGS, is still higher than conventional methods, such as gene panels or microarrays used in CNV detection. However, they are likely to be more cost-efficient overall because they have a higher diagnostic yield, reduce the time to get a diagnosis and may also detect secondary (incidental) findings with clinical relevance (Soden *et al.* 2014; Stavropoulos *et al.* 2016). The clinical utility of WES and WGS is highly dependent on which phenotype is in question. These methods are powerful in, for example, severe neurodevelopmental disorders but their diagnostic yield may be much lower in other severe but apparently genetically more complex conditions such as some forms of immunological disorders (Taylor *et al.* 2015).

2.3.4 Processing of next-generation sequencing data

2.3.4.1 Sequence read processing

An analysis pipeline of NGS data aims to transfer the raw sequence data to a high-quality set of annotated variant calls, while doing quality control checks for the data at multiple steps along the analysis process. Given the huge amount of data produced, its processing requires extensive computational infrastructure and bioinformatics expertise. There is not a single correct way to do the data processing, and because of constant method development in the field, there is now a wide range of public and in-house tools for each step of the pipeline. Broad Institute's Genome Analysis Toolkit (GATK) has become, however, the most widely used software package, owing to its active development, comprehensive documentation and user support.

The principal steps of NGS data analysis pipeline are presented in **Figure 3**. NGS technologies produce tens of millions of short sequence reads per one exome. The first step after base calling, generation and quality control of raw sequence reads is to do alignment, i.e., to determine the most likely position of the sequence reads in the reference genome (Flicek and Birney 2009). The most commonly used methods for short-read sequence alignment include BWA and Bowtie, which are computationally efficient and fast (Langmead *et al.* 2009; Li and Durbin 2009). For some sequence reads, it is difficult or even impossible to determine the correct location because the same or highly similar sequence is present in two or more locations in the genome. These regions include segmental duplications which are genomic sites 1-400 kb in length and share >90% sequence identity (Sharp *et al.* 2005). A mapping

quality score, describing the reliability of the real alignment, is given to each read after alignment. Variant calls in reads with low mapping quality are more likely to be false positives. After alignment, it is important to mark reads that have equal start and end positions. These read duplicates arise typically in polymerase chain reaction amplification of DNA as part of sequencing library preparation (Xu *et al.* 2012). If the duplicate reads are not flagged, it can bias variant detection.

At this point, the quality of the aligned sequencing data can be assessed. In the case of WES, one can for example check what proportion of reads mapped to the targeted region. If it is low, it may indicate errors in the target enrichment process in the laboratory. Moreover, it is important to assess the sequencing coverage across the target regions. Samples with low coverage may be resequenced to produce enough data for adequate variant detection. The desired sequencing coverage depends on the application, with clinical WES in particular aiming to have as little gaps in the exome as possible. Because better coverage equals to higher cost, one needs to balance between the coverage and number of samples that the budget allows to sequence. In typical WES studies, mean sequence coverage is 60-150 reads per base (60-150×), but because the coverage is not uniform in WES, only ~80-95% of target regions are typically covered with 20× (a commonly used threshold for high-confidence variant detection) when having above-mentioned mean coverage values. Next steps in the pipeline recommended by GATK are the realignment of reads near indel sites and recalibration of base quality scores, which both improve quality of the data and allow better quality variant detection.

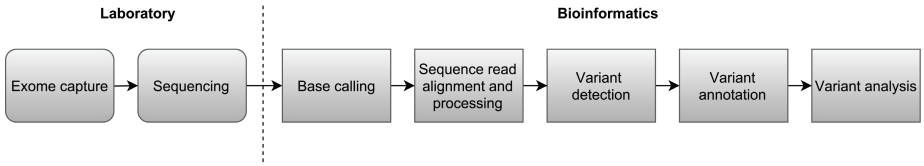


Figure 3 Flow chart of next-generation sequencing data production and bioinformatic processing.

2.3.4.2 Variant detection

Given the importance of obtaining high-quality variant information from sequencing data, variant detection, or variant calling, has been under active method development. Several variant calling tools have been developed to SNV and indel calling, the most commonly used being GATK and samtools (Li *et al.* 2009; McKenna *et al.* 2010; DePristo *et al.* 2011). Methods to detect CNVs from WES data have been discussed by Tan *et al.* (2014). When selecting the variant caller and parameters for variant calling and quality control, one should always test different approaches to obtain best results for own sequencing data. The need for careful assessment of different variant

calling protocols is highlighted by the observed disconcordance between various variant calling pipelines (O'Rawe *et al.* 2013). The most important factors determining the reliability of variant calls is the sequencing depth and quality of base calls. The more high quality sequence reads cover a variant site, the more confident variant calls and genotypes (heterozygous or homozygous) can be obtained. While there has been significant method development, indels are still harder to call reliably compared to SNVs (Fang *et al.* 2014).

There are two main approaches to perform variant calling. The traditional method is to do variant calling separately for each sample. An alternative, more recent approach, which is implement in GATK UnifiedGenotyper and HaplotypeCaller algorithms, is to do variant calling jointly for multiple samples in the same analysis (Liu *et al.* 2013; Lek *et al.* 2015; GATK Best Practices 2016a). The power of the joint approach is that when combining data for each locus from multiple samples, the variant caller has more confidence to determine whether the signal is likely to be a true variant or an artefact (GATK Best Practices 2016a). In other words, the sensitivity to identify low-frequency variant is better and also false positive variants are more confidently detected. Singleton variants, i.e., variants that occur only in one case of the sample set, do not, however, necessarily benefit from the joint calling approach and single-sample calling methods may sometimes provide better sensitivity (GATK Best Practices 2016a). Performing joint calling is important when for example variant data of family members is analysed. When doing single-sample variant calling, incomplete information of the segregation of the variant may be obtained, if, for example, a true variant in one of the individuals is missed because the signal quality happened to be low. Joint analysis would in turn provide genotypes that are more reliable for each member of the family.

Quality-based assessment of generated variant calls is an important task, since raw variant calls contain many false positives. The goal is to do quality-based filtering so that the number of real variants in the filtered data is maximised, while removing as high proportion of artefacts as possible. A traditional method is to set numeric filtering thresholds for several quality-related variant annotation scores, this is so called hard-filtering (GATK Best Practices 2016b). For example, in a very simplified scenario one could filter out variants that have at least one of the following two possible indicators of low quality: 1) strand bias, i.e, alternate allele occurs more often in either positive or negative strand and 2) mapping quality value is low, for example, <20. However, with this method true variants can be missed just because one of the annotations showed indication of low quality even though all the others suggested it is a high-quality variant. On the other hand, to avoid losing true variants one may need to relax the filtering cut-offs, which can lead to increased number of false positives. An alternative approach, which can be used as part of GATK pipeline, aims to bypass these issues by integrating information from the individual quality-based annotations to one

recalibrated quality score (GATK Best Practices 2016b). In this approach, a training set consisting of well-established markers is used to pinpoint likely true variants from the variant call dataset. This way an annotation profile of a good quality variant can be determined. If the annotation profile of a variant differs substantially from those likely to be true, it can be filtered out. With this method, one can select a desired sensitivity and specificity threshold for variant filtering. One can use the newly generated variant quality score to find the appropriate balance between selecting all the possible variants while accepting to have more false positives or selecting only high-confidence variants while risking to lose true variants.

2.3.4.3 Variant annotation

After obtaining a reliable set of variant calls, the initial information about the variant is the genomic position, the nucleotide level change and the genotype. To make use of the data and try to facilitate identification of the causal variant(s) (see section 2.3.5), it is necessary to obtain more information about each variant, including the consequence (missense, synonymous etc.), location (intron or exon, which gene etc.) and frequency (in the general population or disease-specific databases). This process is called variant annotation. Which annotations are needed depends on the application. Widely used annotations in Mendelian genetics are listed in **Table 2**. The list contains both variant- and gene-level annotations that can be utilised in the interpretation of the clinical significance of the variant.

Table 2. Variant- and gene-level annotations for sequence variants.

Annotation type	Example databases	Example annotations
Variant level		
Consequence	VEP, ANNOVAR, SnpEff	Missense, intronic
Gene	Ensembl, CCDS, RefSeq	*SCN1A*
Population allele frequency	ExAC, 1000 genomes, EVS	1.5% in Finns in the ExAC
Clinical variant databases	ClinVar, HGMD	Pathogenic in ClinVar
In-house allele frequency	In-house variant database	30% (a common variant in population or a sequencing artefact)
Evolutionary conservation	phyloP, GERP	Numerical scores (the higher the score, the more conserved site)
In silico deleteriousness prediction	CADD, PolyPhen, SIFT, MutationTaster	Probably damaging (by PolyPhen)
Gene level		
Gene function	Uniprot	Voltage-gated sodium channel
Expression	GTEx	Central nervous system
Associated disease	OMIM	Severe myoclonic epilepsy of infancy

VEP, Variant Effect Predictor; CCDS, Consensus coding sequence, ExAC, Exome Aggregate Consortium; EVS, Exome Variant Server; HGMD, Human Gene Mutation Database; GERP, Genomic Evolutionary Rate Profiling; CADD, Combined Annotation Dependent Depletion; SIFT, Sorting Intolerant from Tolerant; GTEx, Genotype-Tissue expression; OMIM, Online Mendelian Inheritance in Man

There are multiple software packages that can be used to retrieve most of the annotations, one of the most popular ones being Variant Effect Predictor (McLaren *et al.* 2010), ANNOVAR (Wang *et al.* 2010) and SnpEff (Cingolani *et al.* 2012). These tools are user-friendly and typically accept custom annotation sets to be integrated as part of the annotation pipeline. When determining the consequence of variants, an important choice to be made is the gene annotation set based on which the annotation is done. Of the three most widely used gene annotation sets, Ensembl contains the highest number of gene transcripts but it includes many of which biological function is questionable. RefSeq of the National Center for Biotechnology Information (NCBI) is another comprehensive and annotated gene set that is widely used. Consensus coding sequence (CCDS) represents most conservative set of genes, because it represents a consensus gene set of Ensembl and RefSeq. In practice, a single locus may contain multiple overlapping genes and gene transcripts and the annotation of a variant may vary depending on the transcript. For example, at a given genomic position, a SNV may be annotated as intronic but in another transcript the variant encodes a missense variant, because it is located in an exon that is not present in the first transcript. Typically, the transcript having a variant annotation with the most severe predicted consequence is selected while prioritising high quality gene transcripts.

2.3.5 Analysis and interpretation of sequence data to identify disease-causing variants and new disease genes

The appropriate strategy to identify disease-causing variants and novel diseases genes in NGS data depends on severity, mode of inheritance and prevalence of the disease (Gilissen et al. 2012). In fact, these factors should already be taken into consideration when selecting samples that are sequenced. The selection process is naturally affected also by available funds and availability of DNA specimens. Various study and analysis designs for identification of pathogenic variants and novel disease genes from WES or WGS data are presented in **Figure 4**. The whole analysis process is described in the following sections 2.3.5.1-2.3.5.3.

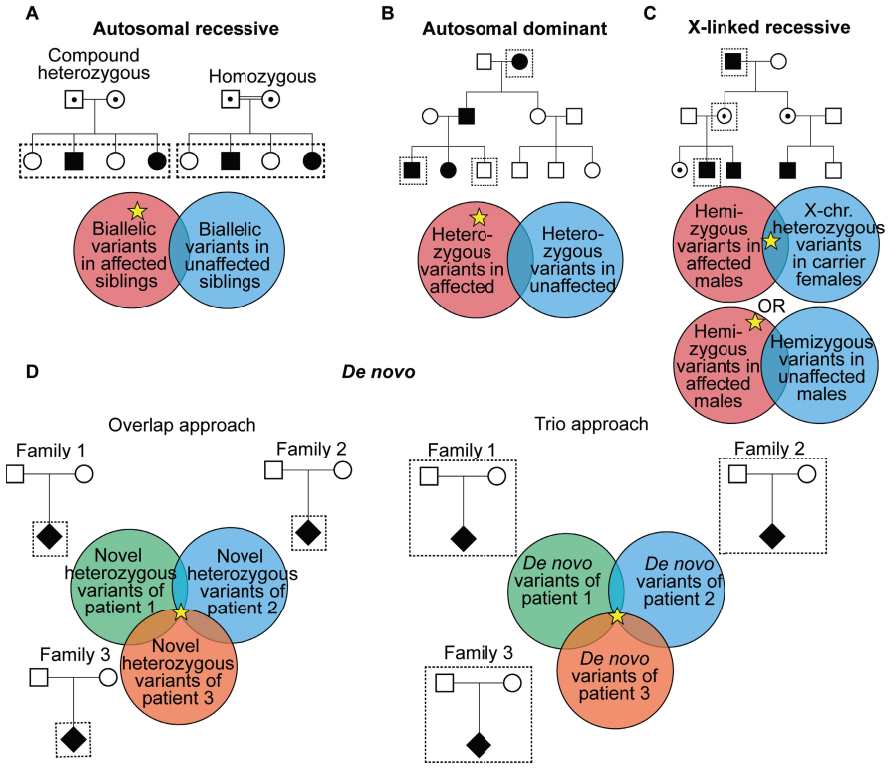


Figure 4 Strategies for identification of disease genes in whole-exome or whole-genome sequencing studies. The figure illustrates suggested sequencing and variant analysis strategies for monogenic disorders with different modes of inheritance (A-D). Dashed lines indicate individuals whose sequencing would provide powerful reduction in the number of candidate variants in the data analysis. Coloured circles below the pedigrees indicate types of variants that are analysed in a disease of given inheritance pattern. Stars indicate the group of variants that are presumed to contain the disease-causing changes. In the panel D is presented two alternative approaches to identify disease genes with *de novo* variants. On the left is presented an overlap approach where the variant data of sporadic index cases are filtered for novel heterozygous variants that are not present in population databases, and the aim is to identify genes where multiple affected individuals have variants. This approach can be successful in disorders with limited genetic heterogeneity and may also be applied with other inheritance patterns. On the right is presented a trio approach, which is similar to the overlap approach but there complete family trios consisting of unaffected parents and an affected child are sequenced to allow direct assessment of *de novo* variants. It is therefore a more powerful strategy. It is also to be noted that in some family trios with sporadic patients the underlying pattern of inheritance may also be recessive, mitochondrial or even autosomal dominant in the case of incomplete penetrance or imprinted genomic loci. Therefore, it is always advisable to use variable strategies in disease gene identification studies. Adapted from Boycott *et al.* (2013) and Gilissen *et al.* (2012).

2.3.5.1 Filtering of whole-exome sequencing variant data

Independent of the utilised NGS application, similar considerations regarding variant filtering apply. To use WES data as an example, exome bait regions, which also contain intronic sequences, contain typically 40,000

variants, while actual exonic regions and exon-intron junctions altogether have 20,000-25,000 variants. The aim of the variant filtering process is to narrow this number down so that identification of the disease-causing variant(s) becomes possible (**Figure 5**). Variant filtering should not be considered as an inflexible process with only one correct way to do it. If filtering parameters used in the initial round of analysis do not yield likely pathogenic variants, one should adjust them, for example, by using other assumptions regarding the mode of inheritance.

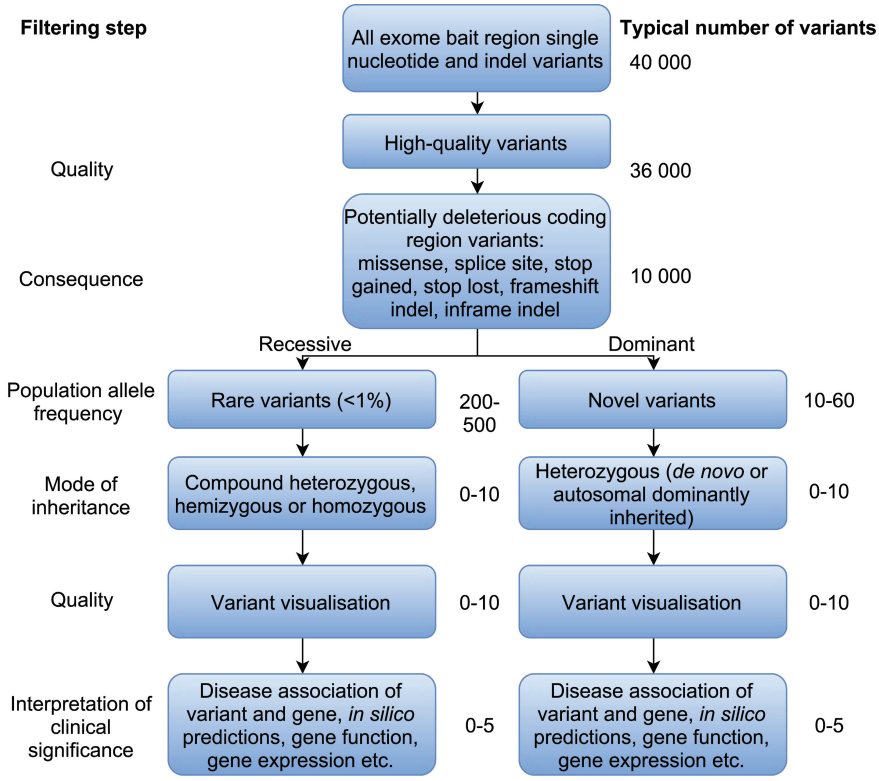


Figure 5 A typical variant filtering process to identify disease-causing variants from whole-exome sequencing data. A filtering protocol is described assuming recessive or dominant/*de novo* inheritance. When filtering for *de novo* variants, the consequence based step can be omitted because the number of resulting variants would be low despite including noncoding and synonymous variants.

One typically included layer in variant filtering is to select changes that are likely to affect the protein sequence. Variants that are usually kept are missense, nonsense and splice site variants as well as inframe or nonframe indels. In this stage, synonymous, intronic and other type of variants outside the coding regions are usually removed. However, while vast majority of these variants are benign, there are several examples of intronic and synonymous variants that are pathogenic because they impair splicing

(Eriksson *et al.* 2003; Wang and Cooper 2007). To assist identification of pathogenic noncoding variants, there are now bioinformatic tools that can be used for any type of variant to predict their deleteriousness and effect on splicing (Desmet *et al.* 2009; Kircher *et al.* 2014; Ionita-Laza *et al.* 2016). However, since these tools are still not sensitive nor specific, the initial round of analysis is usually done to include only coding variants.

A filtering step based on the population frequency of the variants removes the vast majority of the remaining variants. When aiming to identify the genetic cause of a rare Mendelian disease, one can assume that the underlying genetic variants are rare or absent in the general population. Recently, establishment of comprehensive variant databases, such as ExAC, 1000 genomes and Exome Variant Server (EVS), has greatly facilitated annotation of the population frequency of variants. Therefore, identification of variants, which are too common in the population to cause a severe, Mendelian disease, has become more straightforward. These databases contain exome and/or whole genome variant information of adults without severe paediatric diseases. While the databases are a great resource to pinpoint variants that are too common to cause a severe, early-onset disease, one must take into account that some of the individuals in these datasets may have common diseases or will later develop a disease. Therefore, they should not be considered to contain individuals who cannot possess disease-causing variants of any kind, which has implications for variant interpretation (see below). Particular care should be applied when filtering against the NCBI dbSNP database, the most comprehensive genome variation portal, because it may contain disease-causing variants including those submitted originally to the NCBI ClinVar database of clinically-associated variants. When performing variant frequency based filtering, it is important to use frequency information from the appropriate populations. A given variant may appear novel or rare in populations that are included in the commonly used databases, but it may in fact be a common polymorphism in the specific population where a study subject is originated. This is not an issue for individuals of European origin, since those are well represented in reference datasets, but populations from developing countries are generally not well covered.

An appropriate allele frequency threshold used in variant filtering depends on the assumed pattern of inheritance. When filtering variant data under a recessive model, an often-used threshold for allele frequency is 1% but also values higher or lower than that appear in the literature. Selection of frequency cut-offs is done based on the estimated prevalence of the disease but this is often difficult. Therefore, selection of frequency cut-offs is always arbitrary and one must be cautious of not selecting too low values since some of the recessive diseases, such as cystic fibrosis (Cutting 2015), are caused by variants that have relatively high frequencies in the study populations. Particularly in isolated populations, the carrier frequencies of severe Mendelian diseases can be over 1% (Peltonen *et al.* 1999). In dominant

disorders, selecting the frequency threshold is affected by the prevalence and age of onset of the disease as well as the estimated penetrance of the causal variant. If the disease is severe and early-onset, one can initially filter out all variants present in the population-based variant databases or use a very low frequency threshold. However, if the disease is adult-onset and/or variants are not suspected to be fully penetrant, i.e. they do not cause the disease in all carriers, such as in many forms of epilepsies (e.g., Rosanoff and Ottman 2008; Dibbens *et al.* 2013; Ishida *et al.* 2013), a higher value should be used.

After filtering the variant data based on the predicted consequence and the population frequency of the variants, the next step is to include variants that are compatible with the assumed segregation pattern. It is worth noting, however, that variant filtering should not be limited on the most obvious inheritance model(s). For example, an apparently recessive disease may actually be due to heterozygous variants that are germline mosaic in one of the parents. In sporadic cases, both *de novo* and recessive variants should be analysed. In familial cases, either dominant or recessive defects (autosomal or X-linked) are included depending on the segregation pattern of the disease in the family. In consanguineous families, homozygous variants can be prioritised, however, even in these families the pathogenic variants can be *de novo* or compound heterozygous (Powis *et al.* 2016).

Exome sequencing in as many family members as possible, in addition to the index case, facilitates removal of benign variants and identification of the disease causing variants, but it is associated with increasing sequencing costs. In large pedigrees, a cost-efficient solution is to sequence two as distantly related affected individuals as possible and one unaffected relative. Direct detection of *de novo* mutations requires exome sequencing in the parents. An alternative but less powerful approach to identify causal *de novo* variants is to filter variant data for novel heterozygous variants and subject any candidate variants in known disease genes, for example, to segregation analysis in parents. However, if the parental samples are only capillary sequenced for one variant, there is a risk that wrong implications regarding the segregation pattern are made due to sample mix-up or nonpaternity. Therefore, the correct relationship of DNA samples should be checked by, for example, genotyping polymorphic genetic loci.

After filtering variants based on the function and population frequency under assumed inheritance model(s), it is worthwhile to use sequence visualisation tools, such as Integrative Genomics Viewer, to assess the quality of the candidate variants. Despite quality control measures implemented as part of the variant calling process, some artefacts typically remain in the data.

2.3.5.2 Interpretation of variant pathogenicity

Given the complex nature of the genetic variation in the human genome, the assessment of variant pathogenicity is often not a simple task. Giving wrong information regarding the pathogenicity of a variant or missing the likely

genetic cause due to misinterpretation can have dramatic consequences to patient care and counselling. Therefore, interpretation of variant pathogenicity is a process that should ideally utilise several lines of information including genetic, segregation, computational, clinical and functional data (Richards *et al.* 2015).

To improve the quality of variant interpretation and reporting, guidelines for informed and careful assessment of genetic information in clinical setting have recently been suggested (MacArthur *et al.* 2014; Richards *et al.* 2015). American College of Medical Genetics and Genomics (ACMG) has established a five-tier scale to describe the pathogenicity of a variant: pathogenic, likely pathogenic, variant of unknown significance, likely benign and benign (Richards *et al.* 2015). Each of these categories is accompanied with specific criteria that a variant needs to fulfil. These guidelines have become widely adopted in clinical laboratories and some of the general considerations underlying them are covered here.

In the simplest scenario in terms of variant interpretation, the sequence data of a patient may reveal a variant that has been reported to cause a disease in several unrelated individuals in independent publications. If the phenotype of the patient and segregation pattern of the variant are compatible with those linked to the gene previously, it is straightforward to report the variant as pathogenic. A more complex situation occurs when a patient is, for example, compound heterozygous for two rare missense changes that have not been reported in other patients. If the phenotype of the patient has some overlap with that reported previously in patients with same disease gene, it is tempting to classify the variants pathogenic or likely pathogenic. Considering that there are, for instance, hundreds of epilepsy-associated genes, one could observe variants just by chance in one of these genes. Therefore, in this kind of scenario more evidence to support the pathogenicity classification should be obtained. Further support of pathogenicity may be obtained from segregation data but in the case of small families, this evidence alone is not very strong. To estimate the deleteriousness of missense variants, one can utilise computational tools that take into account protein structure and domain information as well as evolutionary conservation of the variant site. Evolutionarily more conserved loci are more likely to be functionally important and that is why these residues are generally enriched for disease-causing variation (Cooper *et al.* 2010b). However, because the accuracy of these *in silico* prediction tools is currently not ideal, they should be used with caution (Hicks *et al.* 2011). Functional laboratory assays may give insight on the effect of a variant on protein function, but for many genes of unknown function or for those where an assay has not been developed, this is not a possibility. Furthermore, even if an assay exists, it may be costly and laborious, thus hindering its utility in everyday clinical testing. Hence, development of high-throughput methods to perform functional assessment of a large number of missense variants is a major area for further research (Starita *et al.* 2015). It also important to bear

in mind that not all variants with a functional effect are pathogenic (MacArthur *et al.* 2014), because, for example, a cell may tolerate reduced or even completely abolished gene function. Therefore, results from functional assays should be analysed in relation to those variants known to be both functional and pathogenic (e.g. reduction in enzymatic activity at least 40%) and to those that are benign (e.g. less than 40% reduction).

With the emergence of large-scale population variant databases and increased knowledge regarding variant interpretation it has become evident that many variants initially reported as pathogenic and included in databases of clinically-associated variants are actually likely to be benign (MacArthur *et al.* 2014; Lek *et al.* 2015). In the early years of genetic research the number of control samples used was much lower, and now, data in the recently established comprehensive population variant databases indicate that many variants initially interpreted as pathogenic are actually too prevalent to cause, for example, a rare dominantly inherited disease (Minikel *et al.* 2016). Along the same line, because information in the disease variant databases, such as ClinVar and Human Gene Mutation Database, is not always curated using clinical standards, the pathogenicity of variants should not be assessed only based on information retrieved from these resources (Rehm *et al.* 2013). Furthermore, some of the variants that have turned out to be present also in healthy individuals may be truly causal but they are just not fully penetrant, which poses a major challenge for variant interpretation and reporting results to patients and their family members (Minikel *et al.* 2016).

The fact that each human exome and genome contains a high amount of variation means that the genetic code of each individual has the potential to be made to fit a story that would explain the phenotype in question. This phenomenon has been termed as 'narrative potential' of the human genome (Goldstein *et al.* 2013; MacArthur *et al.* 2014), and highlights the importance of common guidelines for variant interpretation. As with many other clinical tests, perfect sensitivity and specificity is difficult to reach in genetic testing, but with multilevel and careful assessment of variant pathogenicity, risk of misinterpretation is minimised. A recent study showed high disconcordance between variant classifications reported by independent genetic laboratories, which demonstrates that the challenges regarding variant interpretation really need to be addressed (Van Driest *et al.* 2016).

Because Mendelian diseases are individually rare and even more so are the individual variants underlying them, it is therefore of high importance to share clinical and genetic information of patients that are studied in clinical genetics laboratories across the world. This assists re-evaluation of the pathogenicity of variants of which clinical importance has been initially unknown. Identification of additional patients with same variants and same symptoms would indicate that the variant is indeed pathogenic. Several resources aiming to improve data sharing and our understanding of clinical relevance of genetic variants have been established. Web platforms enabling sharing of genetic and/or clinical data include ClinVar, DECIPHER, LOVD,

and MatchMaker Exchange (Firth *et al.* 2009; Fokkema *et al.* 2011; Philippakis *et al.* 2015; Landrum *et al.* 2016). These efforts are valuable not only in the interpretation of variants of a single patient but they are also designed to match 'siloed' unsolved patient exomes and genomes across various laboratories to facilitate identification of new disease genes.

2.3.5.3 Identification of new disease genes from whole-exome or whole-genome sequencing data

When WES or WGS is performed in a research setting, the main goal is usually to identify new disease genes. The basic principle is to identify genes that are recurrently mutated in the patient cohort after filtering the variant data under one of the inheritance models. When for example *de novo* variants in a single novel disease gene are identified in 12 individuals in a phenotypically homogeneous cohort of 15 cases, the establishment of a new disease gene is already quite evident based on that observation alone. In less extreme scenarios with the disease gene present in the minority of the patient cohort, further evidence to support identification of a new disease gene may be obtained, as discussed above, using segregation analysis and functional assays in cell and animal models, which can confirm with high confidence the pathogenic role of the newly identified variants.

Identification of new disease genes in WES or WGS cohorts can be guided by statistical approaches that pinpoint genes, which possess more variants in cases than expected. Statistical methods are needed because one can observe recurrently mutated genes in large patient cohorts just by chance. For example, pathogenicity of newly identified variants is highly uncertain if *de novo* variants in the same gene are observed in two individuals in a cohort of 500. On average, each individual has ~0.8 *de novo* mutations that alter the protein coding sequence, so in a large WES or WGS sample set this kind of event could happen just by chance, in particular in genes with a lengthy protein coding region (Kong *et al.* 2012; MacArthur *et al.* 2014). To test if there is an enrichment of *de novo* mutations in certain genes in a patient cohort, it is now relatively well-established to use statistical approaches to demonstrate that in the study sample, damaging *de novo* mutations (missense, loss-of-function) occur more often in a given gene than is expected (Samocha *et al.* 2014). These statistical frameworks take into account the length and sequence structure of the gene and estimate what is the baseline mutation rate for a given gene in the patient cohort. If the number of observed *de novo* mutations is significantly higher than the expected value, it implies a disease-causing role for the gene. Naturally, functional studies and/or replication in additional cases are ideally used to confirm the pathogenic role of the mutations. This statistical approach has been successful in pinpointing new disease genes in large-scale sequencing efforts in parent-offspring trios of various genetic disorders (e.g., Epi4K Consortium and Epilepsy Phenome/Genome Project 2013; EuroEPINOMICS-RES Consortium *et al.* 2014; The Deciphering

Developmental Disorders Study 2015). It is a more challenging task to establish a computational framework to identify recessive disease genes in WES cohorts primarily consisting of small families with sporadic patients. Recently, however, Akawi and colleagues utilised a statistical approach to characterise new disease genes where unrelated affected offspring have potentially deleterious, recessively inherited variants more often than expected (Akawi *et al.* 2015). This statistical assessment of the likelihood of the observed genotypes was complemented with evaluation of the likelihood of the phenotypic similarity between cases with variants in the same gene.

The extreme rarity of some genetic disorders may make it practically impossible to find definite statistical support to back up newly identified suggestive genotype-phenotype correlations, as noted by MacArthur and others. While identification of independent families with variants in the same gene is the general requirement for establishment of a new disease gene, it is valuable for other researchers to share genetic, clinical and functional data in publications even if the initial evidence is only from single families. When reporting suggestive gene associations, all supporting data must be reported as comprehensively as possible while being open and realistic about the extent of confidence in causality. (MacArthur *et al.* 2014)

Taken together, NGS technologies coupled with recent analytical developments have allowed more efficient gene discovery. As opposed to traditional methods, it is now possible to identify new disease genes even in the most challenging scenarios, such as in sporadic cases and sample collections with substantial genetic and clinical heterogeneity. Highlighting the increasing pace of gene discovery during the recent years, 30% of solved cases in a study conducted during 2012-2014 had variants in genes discovered in year 2011 or later (Yang *et al.* 2014). It has been estimated that most of the remaining ~3000 unsolved Mendelian diseases are likely to be solved by the year 2020, if not even earlier (Boycott *et al.* 2013).

2.4 Epilepsies

2.4.1 Definition and classification

Epilepsies are diseases that involve epileptic seizures, which can be defined as "a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain" (Fisher *et al.* 2005). According to the latest clinical definition suggested by the International League Against Epilepsy (ILAE), epilepsy is a disease defined by any of the following conditions (Fisher *et al.* 2014):

"1. At least two unprovoked (or reflex) seizures occurring >24 h apart
2. One unprovoked (or reflex) seizure and a probability of further seizures similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years

3. Diagnosis of an epilepsy syndrome"

Epilepsy is then considered as a disease where patients have an underlying pathological predisposition to develop recurrent epileptic seizures. Having a seizure associated with a temporary reduction in the threshold to have seizures does not lead to a diagnosis of epilepsy (Fisher *et al.* 2014). Factors having the ability to lower the threshold and trigger seizures include for example concussion and fever.

Approximately 4% of individuals will develop epilepsy during their lifetime, making epilepsies one of the most common neurological conditions (Hesdorffer *et al.* 2011). Given the relatively high prevalence and adverse effects on quality of life, epilepsies represent a substantial health and economic burden to the patients and the society (Cardarelli and Smith 2010). In addition to adverse health effects of seizures, individuals with epilepsy suffer from stigmatisation in the society (de Boer *et al.* 2008). When considering the global burden of epilepsies, it is also to be noted that the vast majority of individuals with epilepsy are living in developing countries, where access to specialised care and antiepileptic medication are greatly limited (de Boer *et al.* 2008).

Epilepsies are not a single clinical entity but should rather be considered as a group of conditions with seizures being the unifying symptom. Epilepsies can be classified based on a number of criteria including the age of onset, seizure type, origin of the seizure and aetiology (Berg *et al.* 2010). Seizures fall into two main categories: generalised and focal (Berg *et al.* 2010). The former include seizures that involve abnormal brain activity on both brain hemispheres. Generalised seizures can be divided into six further categories: tonic-clonic, absence, myoclonic, clonic, tonic and atonic. Tonic-clonic seizures, for example, initiate with the tonic phase, which is brief and where consciousness is lost, body stiffens and one falls down. The following clonic phase lasts longer and involves primarily rhythmic muscle contractions. In focal seizures, epileptic brain activity is localised to one brain hemisphere and the related symptoms depend on the origin of seizure, given that various brain regions have different functions. One may for example stay conscious during a focal seizure but not always.

According to ILAE, three are three distinctive aetiological classes of epilepsies: genetic, structural/metabolic and unknown (Berg *et al.* 2010). In genetic epilepsies seizures are considered to be caused directly by known or presumed genetic lesions, however, the definition does not rule out the role of environmental factors. This category was formerly called idiopathic, indicating that the underlying cause of the disease is unknown. It is estimated that genetic factors underlie 70% of all epilepsy cases (Hildebrand *et al.* 2013). Genetic generalised epilepsies (GGE) constitute the most common form of genetic epilepsies accounting for approximately 30% of all cases (Panayiotopoulos 2005). GGEs emerge typically in the childhood or adolescence and they are generally not associated with cognitive dysfunction

or developmental delay. Seizures are generally well controlled with appropriate antiepileptic drugs in GGEs (Panayiotopoulos 2005). Epileptic encephalopathies (EE) are another subgroup of genetic epilepsies. They are individually rare but collectively account for an important fraction of epilepsies. EEs are early-onset and severe forms of epilepsies where epileptic brain activity contribute to cognitive and behavioural defects (Berg *et al.* 2010). They do not respond well to antiepileptic drugs and are associated with developmental delay or regression and poor prognosis (Khan and Al Baradie 2012).

Genetic epilepsies can be further divided into epilepsy syndromes with distinctive electroclinical features and these can be arranged based on their onset (neonatal, infancy, childhood, adolescence-adult) and characteristics findings in electroencephalogram (EEG) (Berg *et al.* 2010). EEG is an important diagnostic tool in epilepsies and it records the combined electrical activity of the brain with brainwave patterns (Noachtar and Rémi 2009). These patterns are abnormal in epileptic brain and many epilepsy syndromes have a typical EEG pattern. Magnetic resonance imaging (MRI) is another commonly used method in epilepsy diagnostics, since it can reveal structural defects in the brain that underlie the seizures.

The second aetiological category includes epilepsies that are caused by structural lesions or metabolic conditions (Berg *et al.* 2010). Structural defects, which increase the risk of epileptic activity, may arise due to acquired disorders such as trauma, stroke and infection. Notably, this category can also include genetic disorders, when the underlying genetic variants result in structural lesions or other defects causing seizures. Tuberous sclerosis is one example of such disorders.

It is important to note that symptoms are not limited to seizures in many epileptic syndromes. The effect of the underlying causal factors is not often restricted exclusively on mechanisms regulating seizure activity but also other aspects of development and function of the central nervous system (CNS) and other tissues can be affected. Therefore, individuals with epilepsy may have other neurological symptoms, dysmorphic features, muscular disorders or other defects in any tissue. Seizures are also a common comorbidity of various other neurological and neurodevelopmental disorders where epilepsy is not the primary phenotype. These include, for instance, autism spectrum disorders, migraine and intellectual disability, as well as many other genetic syndromes (Ottman and Lipton 1994; Myers and Mefford 2015).

2.4.2 Genetics of epilepsies

As mentioned in the previous section and illustrated in **Figure 6**, genetic factors underlie majority of epilepsies. Even though epilepsies have been considered a sacred or magical disease throughout the human history, already Hippocratic texts two and half millennia ago postulated the importance of the inherited component in epilepsies (Riggs and Riggs 2005).

The role of genetic factors in epilepsies has been formally demonstrated in family-based studies showing that relatives of affected individuals are in higher risk to have epilepsy and in twin studies showing that monozygotic twins have higher concordance of epilepsy compared to dizygotic twins (Annegers *et al.* 1982; Berkovic *et al.* 1998).

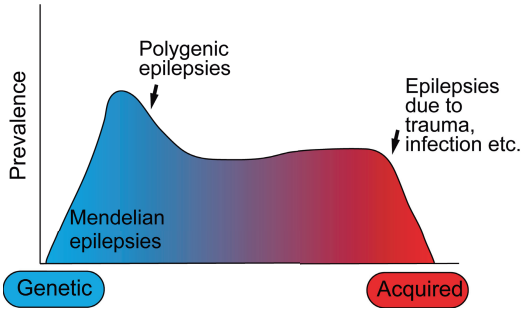


Figure 6 Contribution of genetic factors to epilepsy. The figure illustrates the contribution of genetic and acquired/environmental factors in the aetiology of epilepsies. Genetic factors underlie the majority of epilepsies, the most common forms being polygenic and rare forms monogenic. In addition, acquired forms of epilepsies with little genetic contribution underlie a substantial proportion of epilepsies. Adapted from Hildebrand *et al.* (2013).

First gene discoveries in epilepsies were done in the 1990s. In 1990, the genetic defect underlying myoclonic epilepsy and ragged-red fiber disease, a syndromic form of epilepsy involving myopathy and spasticity in addition to myoclonic seizures, was identified in the mitochondrial genome (Shoffner *et al.* 1990). In epilepsy syndromes where seizures are clearly the predominant clinical feature, the first causal variant was identified in *CHRNA4*, encoding a nicotinic acetylcholine receptor subunit (Steinlein *et al.* 1995). This variant causes autosomal dominant nocturnal frontal lobe epilepsy. One of the most important gene discoveries occurring in the turn of the millennium was identification of pathogenic variants in *KCNQ2* and *SCN1A*, the former encoding a neuronal potassium channel and the latter a neuronal sodium channel (Singh *et al.* 1998; Escayg *et al.* 2000). Variants in these genes, either inherited or *de novo*, have turned out to be common causes of less-severe forms of familial epilepsies with no effect on cognition or development but also of severe forms of EEs (Claes *et al.* 2001; Weckhuysen *et al.* 2012). Despite these individual success stories in epilepsy genetics, in general, as with other genetic disorders, pace of epilepsy gene discovery was relatively slow until the NGS era.

2.4.2.1 Genetics of common and focal epilepsies

The genetic background of most common forms of epilepsies, GGEs, is considered to be primarily oligo- or polygenic with multiple genetic factors and environmental factors contributing to the disease onset (Marini *et al.*

2004; Myers and Mefford 2015). It is therefore not surprising that genes which would individually explain a large proportion of cases have not been identified and that the vast majority of GGE cases remain genetically unexplained. In families where single-gene forms of GGEs have been identified, it is typical that the underlying variants are not fully penetrant and show intra-familial clinical variability ranging from no clinical diagnosis of epilepsy to different manifestations of epilepsies (Meisler MH 2012). This highlights the role of genetic modifiers and other factors in these diseases.

Importantly, over the past few years, CNVs have been implicated in GGEs. For example, recurrent microdeletions in chromosomes 15q11.2, 15q13.3 and 16p13.11 are enriched in individuals with GGE (i.e., they are not fully penetrant), and each variant is found in <1% of all cases (Helbig *et al.* 2009; de Kovel *et al.* 2010). Additionally, microdeletions in other loci overlapping with neurodevelopmentally important genes are enriched in GGE cases (Lal *et al.* 2015). Therefore, collectively CNVs seem to account for a substantial proportion of the genetic component of GGEs. In common epilepsies, genome-wide association studies (GWAS), a commonly used approach in complex disease genetic to identify disease loci (Stranger *et al.* 2011), have had little success due to small sample sizes of the studies with limited statistical power to detect association. In the largest meta-analysis of GGE GWASs, only two loci were identified, one being *SCN1A* (ILAE Consortium on Complex Epilepsies 2014). The same gene was also one of the few significant loci identified in a GWAS of febrile seizures, further demonstrating the role of *SCN1A* as probably the single most important epilepsy gene underlying a variety of epilepsies (Feenstra *et al.* 2014).

While most cases of focal epilepsies remain unsolved, a few genes have been identified. The most important of them is *DEPDC5*, a member of the mTOR pathway (Dibbens *et al.* 2013; Ishida *et al.* 2013). Interestingly, the role of somatic mutations have been demonstrated in focal epilepsies (Baulac *et al.* 2015; Lim *et al.* 2015). It remains to be seen what the contribution of somatic mutations is in other epilepsies.

2.4.2.2 Genetics of rare and severe epilepsy syndromes

Empowered by NGS technologies, WES in particular, gene discovery in infantile- or childhood-onset severe epilepsy syndromes including EEs has accelerated during the past years with tens of new genes identified (reviewed by Myers and Mefford 2015). The success reflects the fact that most of them are Mendelian and caused by variants of large effect (**Figure 6**), and thus NGS provides a powerful platform to dissect their genetic basis. Compared to other genetic disorders, diagnostic yield of WES in severe epilepsies appear to be higher based on a recent study: 38% vs. 28% (Helbig *et al.* 2016). This suggests that interpretation of disease-causing variation is more straightforward in severe epilepsies. Possible causes could be that the genes mutated in severe neurological disorders have less background variation compared to, for example, those in immunological disorders (Petrovski *et al.*

2013) or that the genetic basis of epilepsy syndromes is generally more simple. Another important factor explaining the success in genetics research of rare epilepsies is that the epilepsy community has joined forces to form large sequencing consortia, which greatly increases power to make new gene discoveries (e.g., Epi4K Consortium and Epilepsy Phenome/Genome Project 2013; EuroEPINOMICS-RES Consortium *et al.* 2014). In addition to epilepsy-specific studies, individuals with epileptic syndromes are often included in other large-scale sequencing efforts (Yang *et al.* 2014; The Deciphering Developmental Disorders Study 2015).

The role of *de novo* mutations in epilepsies was first demonstrated with identification of *SCN1A* as the major gene underlying severe myoclonic epilepsy of infancy (also known as Dravet syndrome) (Claes *et al.* 2001). Gene discoveries in the NGS era have confirmed that majority of the pathogenic variants underlying severe Mendelian forms of epilepsies occur *de novo*, which also explains why many of them remained unsolved using traditional methods.

Pathogenic variants associated with most of the recent gene discoveries are point mutations or small-scale indels, with gain-of-function, loss-of-function or dominant-negative effect on protein function (see also section 2.4.3 for the molecular mechanisms of epilepsies). It is also to be noted, however, that in severe forms of epilepsies large CNVs may explain 7-12% of all cases, highlighting the importance of CNV analysis in the diagnostics (Mefford *et al.* 2011; Sisodiya 2015). Identification of CNVs have in some cases facilitated characterization of new epilepsy genes, such as *STXBP1* (Saitsu *et al.* 2008).

Taken together, epilepsies are associated with remarkable genetic heterogeneity, both of locus and allelic type. First, this is illustrated by the large number of genes linked to epilepsies. Based on current knowledge, there are at least 73 genes where variants can cause a disorder classified primarily under epilepsies (EpiPM Consortium 2015). If disorders where seizures are a comorbid feature are also included, there are ~500 such genes (Ran *et al.* 2015b). Secondly, it has become evident that diseases with similar clinical features can be caused by variants in different genes (e.g., Suls *et al.* 2013) and also that variants within a single gene can result in substantially different epilepsy phenotypes or other related neurological conditions (Noebels 2015). In other words, clinical classifications of epilepsies have little overlap with the genetic aetiology of epilepsies. Based on these observations, utility of gene panels in diagnostics of epilepsies can be questioned, since it is often difficult to select a correct set of genes to be tested and because the number of new epilepsy genes grows so rapidly.

2.4.3 Pathomechanisms of epilepsies

One of the basic mechanisms of the nervous system function is that the billions of neurons, which form complex networks, communicate with each other via synaptic connections. This process is called synaptic transmission,

which can be either excitatory or inhibitory. Synapses can be chemical or electrical, of which chemical synapses are most prevalent. (Purves *et al.* 2011)

Dendrites are a branched structure of a neuron, where synapses are located (**Figure 7**). Thousands of synapses, either excitatory or inhibitory, can be present in each dendrite and they can originate from multiple presynaptic neurons. Dendrites are responsible for processing of these synaptic inputs. Postsynaptic potentials are changes in the membrane potential at the synaptic terminal of the neuron that receives the signal. Postsynaptic potential generated by each active synapse is small. These excitatory and inhibitory postsynaptic potentials are, however, added together and if the sum exceeds a threshold, action potential is triggered. Action potential is the electric signal that travels along the axon and sends information to other neurons. In a simplistic example, if the effect of five excitatory postsynaptic potentials is counteracted by five inhibitory potentials, no action potential is fired. In chemical synapses, the neurotransmitter that is released from the presynapse to the postsynapse via the synaptic cleft determines the nature of the synaptic potential. Gamma-aminobutyric acid (GABA) is the most common inhibitory neurotransmitter, while glutamate is an example of a neurotransmitter in excitatory synapses. (Purves *et al.* 2011)

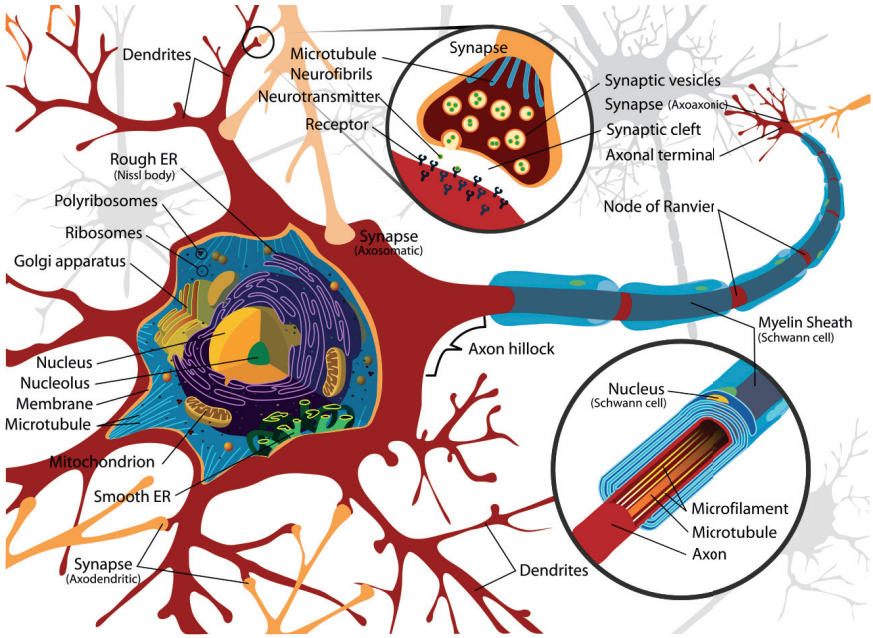


Figure 7 **A neuron.** (The image released into the public domain was reused from https://en.wikipedia.org/wiki/File:Complete_neuron_cell_diagram_en.svg)

A unified pathomechanism for epileptic seizures can be described as the imbalance of inhibition and excitation in the nervous system (Scharfman

2007; Staley 2015). This can result from having too little inhibition (disinhibition) or too much excitation. Thus, seizure activity may arise when the mechanisms inhibiting neuronal firing are impaired or the system facilitating excitation is promoted.

The hundreds of genes known to be linked to epilepsies encode proteins with a large variety functions and many of them do not have any obvious association with potential mechanisms controlling the balance of inhibition and excitation. Initially, however, when first genes associated with epilepsies had been identified, the functional connection of the epilepsy genes and the imbalance between excitation and disinhibition was more evident (George 2004). Many of these genes encode neuronal ion channels, which have an important role in regulation of neuronal excitability. For example, some of these proteins included voltage-gated potassium or sodium channels genes, whose function is to generate and propagate action potentials. Depending on the type of neurons in which the ion channels are expressed and type of the mutation (gain-of-function, loss-of-function), dysfunction of these epilepsy associated ion channel causes either too much excitation or too little inhibition. For example, if a missense variant disrupts sodium channel function in an inhibitory neuron, it leads to decreased inhibition and hyperexcitability of a neuronal network.

Given the strong link of ion channel genes and epilepsies, epilepsies have been considered as 'channelopathies', i.e., diseases where ion channel function is impaired. While many of the newly identified epilepsy genes support the important role of ion channels, the functional spectrum of epilepsy associated genes has expanded dramatically in the NGS era (**Figure 8**). Synaptic transmission, transcriptional regulation, chromatin remodelling and cell growth regulation are examples of pathways that are now linked to severe forms of epilepsies (Myers and Mefford 2015; McTague *et al.* 2016). These proteins can be located anywhere in the neurons, from dendrites to the synapses, and function from early development to further maturation of neurotransmission. Specifically, many of the proteins linked to epilepsies appear to be over-represented in processes controlling neuronal inhibition (Noebels 2015).

Unravelling the genetic basis and molecular pathobiology of epilepsies may provide new targets for antiepileptic treatments. Already now knowing the specific genetic and molecular defect in a patient with epilepsy may guide selecting appropriate treatment options. This works in both ways, a more effective antiepileptic drug that is known to work well with the specific genetic epilepsy syndrome may be selected, and, on the other hand, drugs with known adverse effects may be avoided (Loscher *et al.* 2013). Importantly, recently published studies have provided first evidence that characterisation of the genetic basis of a disease may lead to effective targeted therapies. For instance, inherited or *de novo* variants in *KCNT1*, encoding a voltage-gated potassium channel, cause at least two forms of epilepsies with varying severity (Barcia *et al.* 2012; Heron *et al.* 2012). Some

of the *KCNT1* patients have gain-of-function mutations that cause the ion channel to be too active. Interestingly, functional studies showed that quinidine, which is a known antiarrhythmic molecule, is able to partially reverse the effect of the mutation (Milligan *et al.* 2014). Promisingly, administration of quinidine lead to significant reduction in seizure frequency in two of three patients with *KCNT1* positive epilepsy (Bearden *et al.* 2014; Mikati *et al.* 2015). Other Mendelian epilepsies where genetic diagnosis may assist selection of antiepileptic treatment are those associated with pathogenic variants in *GRIN2A*, *KCNQ2*, *SCN1A* and *DEPDC5* (Baraban *et al.* 2013; Krueger *et al.* 2013; Orhan *et al.* 2014; Pierson *et al.* 2014; Myers and Mefford 2015).

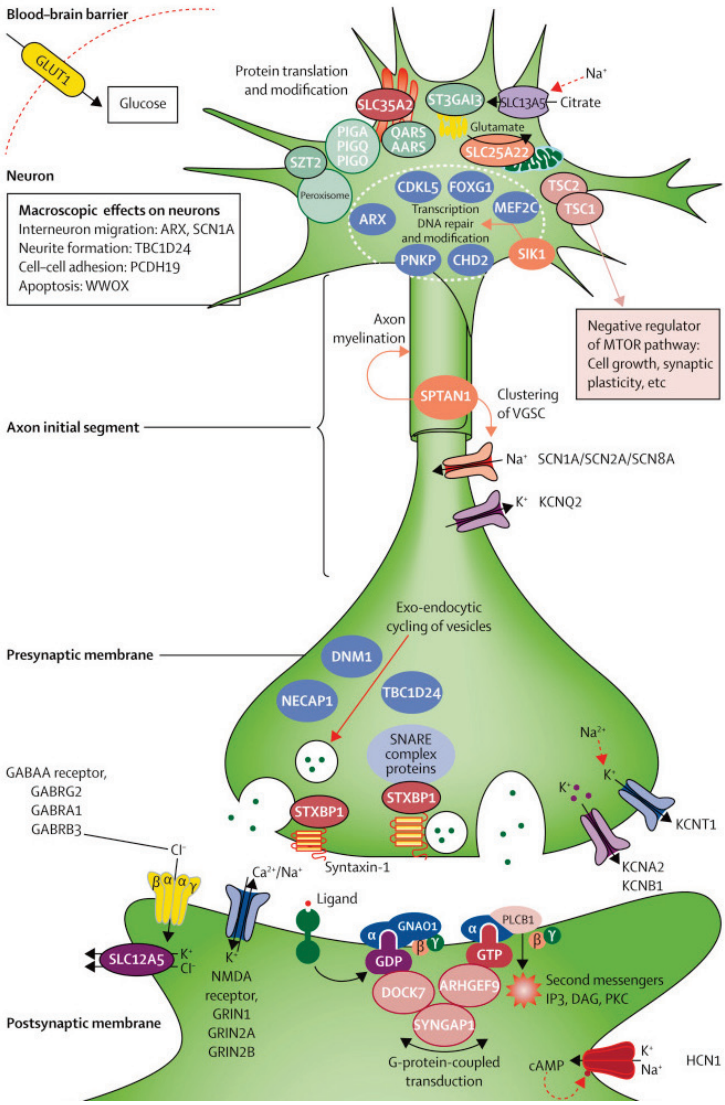


Figure 8 Molecular mechanisms underlying paediatric epileptic encephalopathies. The figure illustrates the wide spectrum of functional roles that proteins defective in epileptic encephalopathies have in the development and function of the central nervous system neurons. Reprinted from The Lancet Neurology, McTague *et al.* (2016), "The genetic landscape of the epileptic encephalopathies of infancy and childhood", © 2016, with permission from Elsevier.

2.4.4 Progressive myoclonus epilepsies

Progressive myoclonus epilepsies (PME), the main focus of this thesis together with severe infantile-onset epileptic syndromes, are a group of rare, genetic and severe epilepsies (reviewed by Berkovic *et al.* 1986; Shahwan *et al.* 2005; Kälviäinen 2015). PMEs are phenotypically heterogeneous but they

are characterised by three main features: myoclonic seizures (brief jerks of a muscle or a muscle group, also called myoclonus, which, notably, is not always a sign of epileptiform activity), tonic-clonic seizures and progressive deterioration of symptoms over time due to neurodegeneration. The disease onset is usually in early to late-childhood or adolescence. In rare cases, PME may also present in adulthood.

In PMEs, myoclonus is typically disabling and affects daily activities. It may occur without stimuli or be triggered by, for example, noise or light. In addition to seizures, PMEs can be associated with other neurological features, such as ataxia (impaired coordination of movements), dysarthria (motor speech disorder) and dementia. Due to disabling myoclonus and ataxia, wheelchair is often needed. In the most common form of PME, Unverricht-Lundborg disease (ULD), cognition is only mildly, if at all, affected and the disease does not affect lifespan. On the contrary, Lafora disease and neuronal ceroid lipofuscinoses (NCL) are examples of PMEs that present with more rapid progression and early death. There is no cure available to any of the PME subtypes and care is symptomatic. Seizures, in particular myoclonus, are generally refractory to medication.

Diagnosis of PMEs in the early course of the disease in particular is challenging due to nonspecific symptoms and clinical overlap with other epilepsies and neurological disorders (Knupp and Wirrell 2014). In some forms of PMEs (Lafora disease, NCLs), inclusion bodies seen in skin biopsies can be used in the diagnostics, but generally, there are not any distinguishing biomarkers. Ultimately, genetic diagnosis is needed for specific diagnosis, and while there are many molecularly defined PME subtypes (**Table 3**), many PME cases are still molecularly unsolved (**Figure 9**) (Franceschetti *et al.* 2014).

The majority of molecularly characterised PMEs are autosomal recessively inherited (**Table 3**). The most extensive study aiming to characterise the genetic aetiology of PMEs was conducted in a nationwide collection of 204 patients in Italy (Franceschetti *et al.* 2014). However, this study screened only known PME genes, and no large-scale efforts dissecting the genetic basis of PMEs using NGS technologies have been conducted prior to this study. Based on the study by Franceschetti and colleagues, ULD (38% of the cohort), Lafora disease (16%) and NCLs (6%) are the most common forms of PMEs (**Figure 9**). However, since these PME subtypes are autosomal recessively inherited, with population specific allele frequencies affecting the disease prevalence, same findings may not apply to all populations. In fact, Western Mediterranean countries together with Finland (1 in 20,000 births) have the highest incidence of ULD worldwide (Shahwan *et al.* 2005).

PME-associated genes encode a functionally diverse set of proteins (**Table 3**), but the pathomechanisms have remained largely uncharacterised. Many of the PMEs are molecularly linked to endosomal or lysosomal dysfunction, which leads to neurodegeneration and seizures. For example,

NCLs are caused by accumulation of lipopigments in lysosomes (Kollmann *et al.* 2013).

Table 3. Genetics of PME subtypes.

PME subtype	Inheritance pattern	Gene(s)	Protein function/molecular pathway
ULD (EPM1)	AR	*CSTB*	Inhibitor of lysosomal cysteine proteases
Lafora disease (EPM2A/B)	AR	*EPM2A, NHLRC1*	Glycogen metabolism
NCLs	AR/AD	*PPT1*, *TPP1*, *CLN3*, *CLN5*, *CLN6*, *MFSD8, CLN8, CTSD, DNAJC5, CTSF, ATP13A2, GRN*	Many lysosomal enzymes or membrane proteins
EPM3	AR	*KCTD7*	Interaction with potassium ion channels?
AMRF (EPM4)	AR	*SCARB2*	Lysosomal membrane protein
North sea PME (EPM6)	AR	*GOSR2*	Golgi vesicle transport
MERRF	Mitochondrial	*MT-TK*, *MT-TF*, *MT-TL1*, *MT-TI*, *MT-TP*	Mitochondrial transfer-RNAs
Sialidoses	AR	*NEU1*	Lysosomal enzyme which breaks down oligosaccharides
DRPLA	AD	*ATN1*	Accumulation of ATN1 in neurons due to repeat expansion
MEAK[1] (EPM7)	AD/*de novo*	*KCNC1*	Neuronal voltage-gated potassium ion channel

PME/EPM, progressive myoclonus epilepsy; ULD, Unverricht-Lundborg disease; AR, autosomal recessive; NCLs, neuronal ceroid lipofuscinoses; AD, autosomal dominant; AMRF, action myoclonus with renal failure; MERRF, myoclonus epilepsy and ragged-red fibers; DRPLA, dentatorubro-pallidoluysian atrophy; MEAK, myoclonus epilepsy and ataxia due to potassium channel mutation.

[1]Identified in this thesis (I).

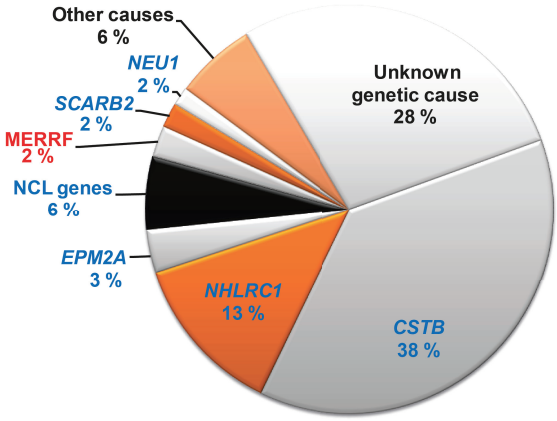


Figure 9 Genetic causes of PME in a nationwide cohort. Data in the figure is from a multicenter study where 204 patients were collected in a 15-year period in Italy (Franceschetti *et al.* 2014). Genes in blue are autosomal recessively inherited and red color indicates myoclonus epilepsy and ragged-red fibers caused by mitochondrial DNA variants.

3 AIMS OF THE STUDY

The aim of this thesis was to decipher the molecular genetic basis of PMEs and severe, infantile-onset epileptic syndromes.

Given the variability in clinical presentations in the studied cohorts, we hypothesised that they are genetically heterogeneous. Moreover, as the study subjects had been screened negative for pathogenic changes in a varying number of known disease genes, we hypothesised that the cohort may be enriched for novel underlying genetic causes. With these considerations in mind, we hypothesised that WES would provide a powerful and hypothesis-free approach to establish new genotype-phenotype associations in these syndromes.

Specifically the aims were:

1. To identify disease-causing genetic variants in individuals with unexplained PME or severe, infantile-onset epileptic syndrome by WES.

2. To identify novel genes underlying these syndromes and to confirm pathogenicity of the identified variants by functional assays.

3. To improve diagnostics of the syndromes and to provide new insight on the underlying biological mechanisms of epilepsies.

4 MATERIALS AND METHODS

4.1 Study subjects

4.1.1 Progressive myoclonus epilepsy patients (I)

The WES cohort consisting of 84 unrelated individuals with PME was collected through international collaboration in over a 25-year period. The collection of subjects and their clinical information was coordinated by Professors Samuel F. Berkovic (University of Melbourne, Australia) and Anna-Elina Lehesjoki (University of Helsinki, Finland). Approximately half of the patients showed a clinical presentation similar to ULD, i.e. onset at late childhood or adolescence and cognition not deteriorated. Majority of the cases (73) were of European descent and approximately half Italian. The remaining cases were of Western Asian (7), Southern Asian (3) or Chinese (1) origin. All individuals were excluded for variants in *CSTB*, the most common PME gene. The extent of other genetic studies varied among the study cohort. Parents and family members were not subjected to WES but were recruited for segregation analysis of identified candidate variants.

Seventy probands had no previous family history of PME. Three cases had either an affected parent or offspring. Family pedigrees of eleven cases were suggestive of an autosomal recessive inheritance of the disease. Fifteen patients were reported to be born to consanguineous parents. Analysis of exome variant data for inbreeding by FEstim (Leutenegger *et al.* 2003) implemented in FSuite v. 1.0.2 (Gazal *et al.* 2014) indicated that eighteen patients were from consanguineous families. Cryptic relatedness between study subjects was not detected by PLINK identical-by-descent analysis (Purcell *et al.* 2007).

Additionally, a secondary cohort of 28 individuals with PME or possible PME was collected for Sanger sequencing screening of candidate variants identified in the primary cohort.

4.1.2 Finnish patients with severe, infantile-onset epileptic syndromes (II,III)

Children showing severe, infantile-onset epileptic syndromes with suspected genetic aetiology were collected in over 20-year period via Finnish university hospitals with the aim to identify the underlying genetic cause. Eventually, 30 children without a genetic diagnosis were subjected to WES. A manuscript of the analysis of the whole WES cohort is under preparation (A. Laari, M. Muona *et al.*), while identification of novel disease genes in families in the WES cohort is described in studies II and III included in this thesis.

Patients in the study cohort have severe neurodevelopmental conditions with epilepsy. A subset of these syndromes may be considered as EEs where seizure activity itself contributes to cognitive and behavioural dysfunction

and disease progression. In the remaining cases, neurodevelopmental defects precede the onset of seizure activity, i.e., genetic defects are not directly linked to seizure development. All cases had been screened negative for the Finnish founder mutation in the gene underlying the PEHO syndrome, which is an early-onset encephalopathy with epilepsy belonging to the Finnish disease heritage (A.-K. Anttonen *et al.* under review). Majority of patients were screened for large CNVs either prior to this study or as part of this study (see section 4.2.10).

In study II, a nonconsanguineous family with two affected siblings was ascertained within the above described sample collection (denoted as "family A"). The patients have an infantile-onset encephalopathy which initially presented with irritability followed by dystonic movements, epilepsy, intellectual disability and stagnation of development. WES was performed for the parents and the index case (II, **Figure 1B**). In study III, a proband with a rapidly progressing encephalopathy with epilepsy was ascertained within the same study cohort and subjected to WES. The healthy parents were not known to be closely related. The proband has two unaffected siblings and no family history of epilepsy (III, **Figure 2A**).

Additionally in study II, a nonconsanguineous family with four affected siblings manifesting with a similar disease to patients in family A in study II was ascertained independently from the above described sample collection (denoted as "family B"). Two affected siblings, one unaffected sibling and parents were subjected to WES (II, **Figure 1B**). A detailed description of the patients' symptoms in families A and B is presented in the original publication II.

4.1.3 Ethical issues (I-III)

Informed consent for DNA analysis was obtained from patients or legal guardians of the patients in line with local institutional review board requirements at time of collection. The genetic studies were approved by the ethical board of Helsinki University Hospital (I-III; decision no. 424/E7/2002, amended in 2011, and 183/13/03/03/2009).

4.2 Production and analysis of genetic data

4.2.1 Whole-exome sequencing (I-III)

WES was carried out at the Wellcome Trust Sanger Institute (WTSI; Hinxton, Cambridge, UK) in all studies with the exception of members of family B in study II that were exome-sequenced at the sequencing core facility of the Institute for Molecular Medicine Finland (FIMM; Helsinki, Finland).

Briefly, genomic DNA extracted from peripheral blood of subjects was fragmented and subjected to library creation using standard Illumina (San Diego, CA, USA) paired-end protocols. For samples sequenced at WTSI,

exome targets were captured with SureSelect Human All Exon 50 Mb V3 RNA baits (Agilent Technologies, Santa Clara, CA, USA). For samples processed at FIMM, Nimblegen SeqCap EZ Human Exome Library v2.0 was used. Paired-end sequencing (reads 75 bp at WTSI, 93 bp at FIMM) was performed using Illumina HiSeq 2000 (WTSI) or HiSeq 1500 (FIMM).

4.2.2 Sequence read processing (I-III)

Sequence reads produced at WTSI were aligned to the human reference genome using the Burrows-Wheeler Alignment Tool v. 0.5.10-mt (Li and Durbin 2009). The 1000 Genomes Phase II reference (hs37d5) was used as the reference genome. It is based on GRCh37 and consists of chromosomes 1–22, X, Y, as well as the mitochondrial genome (rCRS mitochondrial sequence NC_012920). Duplicate reads were marked using the Picard toolkit v. 1.107 (http://broadinstitute.github.io/picard/). Sequence reads were further processed by performing local realignment around known indel locations and recalibrating base quality scores were recalibrated with GATK tool package v. 2.8.1 (McKenna *et al.* 2010; DePristo *et al.* 2011; Van der Auwera *et al.* 2013).

Sequence reads produced at FIMM were processed by the bioinformatics core facility of the institute. Briefly, reads were processed and aligned to the human reference genome hg19 (GRCh37) with a protocol described by Sulonen and colleagues with minor modifications (Sulonen *et al.* 2011).

4.2.3 Single-nucleotide and indel variant calling (I-III)

As a primary approach, sequence variants were called 'jointly' so that all exomes of a dataset were analysed simultaneously. The exomes as part of the joint calling process were sequenced at WTSI, i.e., exomes of individuals from family B in study II were called separately because they were sequenced at FIMM. In the joint approach, SNVs and indels were called using HaplotypeCaller algorithm of GATK (v. 2.8.1 in study I, v. 3.3 in studies II and III). In study I, variants for the 84 PME exomes were called jointly. In studies II and III, the variant calling dataset consisted of the 84 PME exomes accompanied with 43 exomes from the infantile-onset epileptic syndrome cohort described in section 4.1.2. The 43 exomes from the infantile onset epileptic syndrome cohort includes thirty patient exomes, ten parental exomes and three exomes from other samples. Recalibration of variant quality scores and quality-based filtering of called variants was done with GATK VariantRecalibrator as per GATK Best Practices protocol with minor modifications. In study I, a sensitivity cut-off of 99.0% was used for both SNVs and indels. In studies II and III, this cut-off was 99.75%. In addition, individual variant calls with less than five reads at the site were filtered out. We also utilised single-sample variant calling methods (GATK, samtools) to compare the performance of different approaches and to confirm that any variants in known disease genes or potential new genetic causes were not

missed using the joint approach. For the comparison variant call sets obtained with single-sample calling methods, quality-based assessment of variants was done using 'hard filters' on various quality annotation scores.

For exomes generated at FIMM (study II, family B), SNVs and indels were called using samtools (Li *et al.* 2009) as described previously (Sulonen *et al.* 2011).

4.2.4 Analysis of mitochondrial DNA variants (I,III)

Variants in mitochondrial DNA (mtDNA) underlie one form of PME (Myoclonic epilepsy with ragged-red fibers) (Shoffner *et al.* 1990) and early-onset syndromes with seizures (Mitochondrial encephalopathy with lactic acidosis and stroke-like episodes and Neuropathy, ataxia, and retinitis pigmentosa*)* (Goto *et al.* 1990; Holt *et al.* 1990). Hence, sequence data were analysed for pathogenic changes in mtDNA in studies I and III where mitochondrial inheritance was a possibility. Exome capture kits used in this study did not contain baits for mtDNA, but due to abundance of mtDNA in cells, enough mtDNA sequence reads were produced to allow detection of variants with a moderate heteroplasmy level (average read depth of 32.7× in study I and 35.7× in study III). Variants were called using GATK UnifiedGenotyper (study I) or samtools (study III; Li *et al.* (2009)). Known mtDNA polymorphisms and pathogenic variants were annotated based on the MITOMAP database (http://www.mitomap.org).

4.2.5 Variant and gene annotation (I-III)

Variant consequences were annotated using Variant Effect Predictor (release 75 in study I, release 78 in studies II and III; McLaren *et al.* (2010)). *In silico* prediction of deleteriousness of missense variants was done using CADD (Kircher *et al.* 2014), PolyPhen HumVar (Adzhubei *et al.* 2010), SIFT (Kumar *et al.* 2009) and MutationTaster (Schwarz *et al.* 2010; Schwarz *et al.* 2014).

Variant frequencies in the general population were assessed based on the following databases: 1000 genomes project (study I: 1092 samples of the phase 1 release; studies II and III: 2535 samples of the phase 3 release), EVS of the NHLBI GO Exome Sequencing Project (studies I-III: 6503 samples of the 0.0.25 release), Finnish exomes of the Sequencing Initiative Suomi (SISu) project (study I: 3268 samples) and ExAC (studies II and III: 60,706 samples of the v. 0.3 release). Notably, ExAC v. 0.3 release contains over 3,200 exomes of the SISu project as well as samples from the 1000 genomes and EVS projects.

Association of individual variants to human disease was annotated based on NCBI ClinVar database (http://www.ncbi.nlm.nih.gov/clinvar/), gene-specific online variant databases and literature search. Association of genes to human disease was annotated based on the Online Mendelian Inheritance in Man, OMIM® database (Hamosh *et al.* 2005), neurological disease gene panels (Lemke *et al.* 2012) and literature search.

4.2.6 Variant filtering and analysis (I-III)

Variant data were initially analysed for changes in previously established disease genes. For individuals who were without likely pathogenic variants in known disease genes, variants in novel disease genes were explored.

We applied variant filtering strategies based on the possible underlying inheritance patterns (**Figure 10**). As described in section 4.1, the majority (70) of individuals with PME in study I are sporadic so the exome data was analysed for recessive (autosomal and X-linked), *de novo* (an indirect approach was used since parents were not exome-sequenced, see also below) and mtDNA variants (I, **Figure 1a**). In 11 'recessive' families (probands have affected siblings or cousins) and three 'dominant' families (affected parent-offspring pair), recessive or dominant filtering strategy was primarily applied, respectively. In study II, where the suggestive pattern of inheritance was autosomal recessive in both families A and B, autosomal recessive variants were primarily searched while considering the possibility of germline mosaicism of heterozygous variants in either parent (II, **Table S2**). In study III, where the exome of a sporadic female patient was analysed, autosomal recessive, *de novo* (indirect approach) and mtDNA were analysed (III, **supplemental figure e-1**).

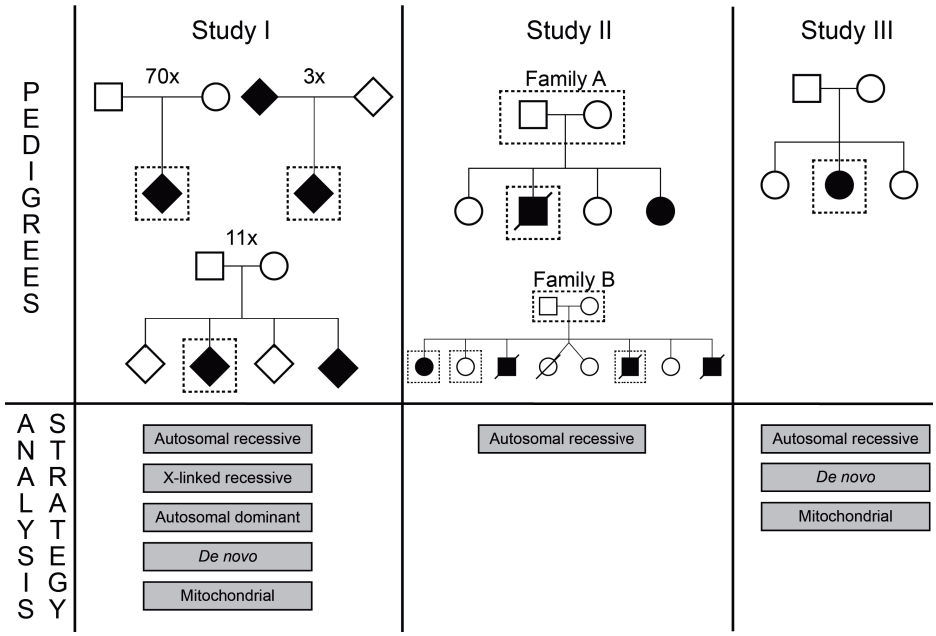


Figure 10 Pedigree structures and primary variant filtering strategies of studies I-III. Dashed lines indicate individuals who where whole-exome sequenced. Illustrations of pedigrees of studies II and III are actual representations of the studied families, whereas the pedigree pictures of study I are simplified and only indicate the various types of family structures that are present in the study.

Variants within gene transcripts of the CCDS database, implemented in Ensembl release 75 (study I) or 78 (studies II and III), were primarily considered but less conservative Ensembl gene transcripts were also used in parallel. Variants with the following consequences were included: missense variant, initiator codon variant, splice donor or acceptor variant, stop lost, stop gained, inframe insertion or deletion, and frameshift variant. Inframe indels in tandem repeat regions (Benson 1999) obtained from UCSC Genome Browser (Kent *et al.* 2002) were excluded.

In recessive filtering, homozygous, hemizygous (when X-linked variants were analysed) and potentially compound heterozygous variants were included. Variants with <1% allele frequency and those not present as homozygous or hemizygous in the databases listed in section 4.2.5 were included. Since parents were not exome sequenced (besides in study II), phase information of heterozygous variants (including compound heterozygous) could not be directly assessed except for those residing in same sequence reads. As an indirect approach, we utilised phase information of variants from the 1000 genomes and SISu data. Ultimately, the phase of candidate compound heterozygous variants was done by analysing the segregation with Sanger sequencing (see section 4.2.7).

In studies I and III, we analysed *de novo* variants using an indirect approach, since parental samples were not exome sequenced. We included only heterozygous variants that are not present in the population variant databases listed in section 4.2.5 or in NCBI dbSNP database (build 138; http://www.ncbi.nlm.nih.gov/snp), except clinically associated variants in NCBI ClinVar. In other words, we assumed complete penetrance of the variants. This was done because in study I, most patients have disease onset before adulthood with maximum onset at 26 years and in study III, patient had a severe early-onset disease. In study III, all variants passing the '*de novo*' filtering were subjected to segregation analysis by Sanger sequencing, whereas in study I only selected variants were further analysed (see sections 4.2.8 and 4.2.9).

Quality of variants passing the filtering was assessed with Integrative Genomic Viewer (https://www.broadinstitute.org/igv/; Robinson *et al.* (2011)). Majority of low quality variant calls were already flagged at steps described in section 4.2.3, but some false positive variants, many of which occur in segmental duplication regions, were removed at this stage. The final high-confidence variants were analysed for pathogenic changes (see sections 4.2.8 and 4.2.9).

Additional details regarding the variant filtering process are presented in original publications of the studies.

4.2.7 Variant validation and segregation analysis (I-III)

Candidate variants identified by WES were validated and their segregation analysed by bidirectional Sanger sequencing using ABI BigDye 3.1 chemistry and ABI 3730xl DNA Analyzer (FIMM sequencing laboratory, Helsinki,

Finland). Primers were designed using Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/; Ye *et al.* (2012)) and sequences analysed with Sequencher (Gene Codes Corporation, Ann Arbor, MI, USA).

4.2.8 Variant pathogenicity classification (I)

In study I, interpretation of variants was done according to similar considerations as recently proposed by ACMG (Richards *et al.* 2015). Clinicians were involved in the interpretation process for each case.

We classified variants passing the filtering as pathogenic, probably pathogenic or unlikely pathogenic. Detailed classification criteria are presented in the original publication of study I. Briefly, factors we took into consideration were among others compatibility of the patient phenotype to those in previously reported cases, previous reports of the pathogenicity of the variant, conservation of the altered amino acid and information of the functional domains of the protein. To be classified as pathogenic or probably pathogenic, segregation pattern of the variant(s) had to concur with that of reported for previous cases in the literature.

4.2.9 Search for novel progessive myoclonus epilepsy genes in the exome variant data (I)

PME exomes, where analysis of known disease genes did not yield probably pathogenic or pathogenic variants, were collectively analysed for variants in novel disease genes (n=69). We used overlap based strategy (**Figure 4**, **p. 36**), i.e., we looked for genes with variants in multiple PME exomes. After filtering variants based on recessive or dominant/*de novo* approaches, we ranked the remaining genes based on the number of cases with variants in the gene. As novel candidate genes for PME, we considered genes with two or more cases with variants passing the filtering in the recessive approach. In the dominant/*de novo* approach, we used a threshold of four genes. We used the higher threshold, since not having parental exomes lead to a substantially higher number of candidate genes passing the dominant/*de novo* filtering, making the analysis of genes with low patient count unfeasible.

In the prioritisation of candidate genes for follow-up, we considered the following factors. We gave less priority to highly polymorphic genes in the general population using a score evaluating the tolerance of genes for functional variation (Petrovski *et al.* 2013) and applying a list of hyperpolymorphic genes identified by others (Fuentes Fajardo *et al.* 2012). We also utilised information on the gene function, which was annotated by literature search and the Uniprot database (http://www.uniprot.org/; The UniProt Consortium (2014)), and expression pattern (GTEx, http://www.gtexportal.org/home/; Lonsdale *et al.* (2013)). Finally, to establish the causal role of the gene in the disease, we performed segregation

analysis of variants in available family members, and functional assays (see section 4.3).

4.2.10 Analysis of copy number variants from genome-wide single-nucleotide polymorphism data (III)

In study III, Illumina Human CoreExome single-nucleotide polymorphism array with 548,000 markers was used in the family trio to detect CNVs from genomic DNA isolated from peripheral blood of the proband. PennCNV was used to call CNVs (http://www.openbioinformatics.org/penncnv/; Wang *et al.* (2007)).

4.3 Functional studies of identified variants (I-III)

We collaborated with a research group in each study to assess the pathogenicity of the identified candidate variants in three different genes: *KCNC1* (study I: H. Lerche lab, University of Tübingen, Germany), *UBA5* (study II: M. Komatsu lab, Niigata University, Japan) and *ADAM22* (study III, M. Fukata lab, National Institute for Physiological Sciences, Okazaki, Japan).

As an exception, in study II the author performed analysis of RNA level consequences of *UBA5* variants in two patients (A-4 and B-3) with biallelic *UBA5* variants. Briefly, total RNA was extracted from primary skin fibroblasts of the patients. RNA was converted to complementary DNA (cDNA) and reverse transcription polymerase chain reaction (RT-PCR) was used to amplify cDNA at mutation sites. PCR products were subjected Sanger sequencing to determine variant genotypes in the RNA level. Additionally in patient A-4, *UBA5* cDNA was analysed for splicing defects. RT-PCR with primers targeted to exons 1 and 3 were used to analyse the effect of c.164G>A on exon 2 splicing. PCR products extracted from agarose gel were subjected to Sanger sequencing.

Detailed methodological descriptions of the methods used in the functional analysis are presented in the original publications I-III.

5 RESULTS AND DISCUSSION

5.1 Identification of a new major subtype of progressive myoclonus epilepsies and broadening the genetic and phenotypic spectrum of known disease genes (I)

5.1.1 A likely genetic diagnosis was reached in 31% of cases

WES produced on average 4.15 gigabases of sequence per patient in exonic target regions, corresponding to an average sequencing coverage of 81 reads per nucleotide. On average 92% of targeted regions were captured with at least ten reads, which allowed a solid basis for variant calling in most regions.

We analysed the exome data under recessive and dominant/*de novo* inheritance models and categorised candidate variants using a three-tier classification scheme where likely and confidently disease-causing variants were annotated as probably pathogenic and pathogenic, respectively. Altogether, we molecularly solved 26 out of 84 cases in the cohort (31.0%). Of the solved patients, 15 had variants in previously established disease genes (see section 5.1.2), and 11 in a gene where variants have not previously been linked to human disease (see section 5.1.3).

The vast majority of previously molecularly diagnosed PME cases have autosomal recessively inherited variants (Franceschetti *et al.* 2014). Surprisingly, more than half of the solved cases in our study were due to *de novo* or autosomal dominantly inherited variants (12 *de novo*, 2 autosomal dominantly inherited and 12 autosomal recessively inherited; I, **Figure 1b**). No likely pathogenic variants were identified in mtDNA. The findings are presented in detail in the following sections 5.1.1-5.1.5 and in the original publication (I).

5.1.2 Variants in previously established disease genes expand the genotypic and phenotypic spectrum of PMEs

We first analysed the exome sequencing data for recessive and dominant/*de novo* variants in previously established PME, epilepsy and neurodegenerative disease genes. Using the variant classification scheme similar to that proposed by ACMG (Richards *et al.* 2015), variants in 12 cases were interpreted as pathogenic and in three cases as probably pathogenic (**Table 4**; see also I, **Table 2**). Of these cases, 12 had autosomal recessive, two *de novo* and one autosomal dominant variants. Clinical details, pedigrees and segregation data as well as conservation of the new variants are presented in the supplementary data of the original publication.

Identification of variants in known disease genes expands both the clinical and genotypic spectra of PMEs, and highlights the utility of WES as a

diagnostic tool in a heterogeneous sample collection. Out of the total 17 variants interpreted either pathogenic or probably pathogenic, seven were novel disease variants. Of the fifteen cases with pathogenic variants in known disease genes, ten had variants in genes specifically linked to PMEs. Notably, seven of these cases had an atypical clinical presentation. For example, all three cases with compound heterozygous variants in *NEU1* did not have all or any of the signs of retinal impairment considered as a key clinical marker for NEU1 deficiency (sialidosis). Indeed, recently there has been a report of other similar cases who do not have all clinical signs considered to be pathognomonic for sialidosis (Canafoglia *et al.* 2014). Other examples of atypical presentations are two of the three Lafora disease cases, who present with a milder phenotype than is typically reported. One of the cases is still alive at the age of 42, while Lafora disease patients generally die before the age of 25 after deteriorating rapidly (Jara-Prado *et al.* 2014).

We identified pathogenic or probably pathogenic variants in three known disease-related genes, where PME has not been considered as part of the clinical spectrum. One case had a previously reported pathogenic *PRNP* variant which causes Gerstmann-Sträussler-Scheinker disease, one of the inherited prion diseases characterized by progressive adult-onset ataxia and dementia (Hsiao *et al.* 1989). The variant was likely inherited from similarly affected and deceased father, from who we did not have DNA for segregation analysis. While PME has not been associated with chronic prion diseases, seizures, including myoclonic seizures, have been reported cases with Creutzfeldt-Jakob disease and Gerstmann-Sträussler-Scheinker disease (Brown *et al.* 1986; Young *et al.* 1997; Bianca *et al.* 2003). In two cases, we identified compound heterozygous variants in *SACS* that we interpreted as probably pathogenic. Autosomal recessive variants in *SACS* cause childhood-onset spastic ataxia, initially described clinically relatively homogenous. However, atypical and late-onset forms of the disease, including reports of seizures, have recently been described (Baets *et al.* 2010; Synofzik *et al.* 2013). Two of the three *SACS* variants identified in this study have been previously reported in individuals with ataxia. Finally, in one patient, we interpreted a homozygous missense variant in *TBC1D24* as probably pathogenic. A remarkable degree of clinical heterogeneity has been associated with biallelic variants in *TBC1D24*. Disorders linked to the gene include various forms of epilepsies including infantile myoclonic epilepsy, a multi-organ syndrome DOORS, and nonsyndromic deafness (Corbett *et al.* 2010; Falace *et al.* 2010; Afawi *et al.* 2013; Guven and Tolun 2013; Milh *et al.* 2013; Campeau *et al.* 2014; Rehman *et al.* 2014). Our case, who has brain imaging findings similar to other *TBC1D24* cases, suggests that PME may be added to the list of *TBC1D24* associated disorders.

Finally, we had one individual where WES helped to correct the initial clinical diagnosis. The patient had a heterozygous, previously reported missense variant in *SCN1A* (Harkin *et al.* 2007), where *de novo* mutations cause severe myoclonic epilepsy of infancy (Dravet syndrome). Initially,

symptoms of the patient had suggested PME as the clinical diagnosis, but evaluation of the further clinical course confirmed this case is indeed presenting with Dravet syndrome.

Table 4. Previously established disease genes with pathogenic or probably pathogenic variants in the WES cohort.

Gene	Patients	Patients with atypical symptoms[1]	Patients with new variants	Inheritance pattern	Gene previously linked to PME	Disease linked to the gene
NEU1	3	3	1	AR	Yes	PME (sialidosis)
NHLRC1	2	1	1	AR	Yes	PME (Lafora disease)
AFG3L2	2	2	2	AR	Yes	Ataxia/Spastic ataxia with PME
SACS	2[2]	NA	1	AR	No	Spastic ataxia
EPM2A	1	1	1	AR	Yes	PME (Lafora disease)
CLN6	1	0	1	AR	Yes	PME (Neuronal ceroid lipofuscinosis)
SERPINI1	1	0	-	*De novo*	Yes	PME with dementia
TBC1D24	1[2]	NA	1	AR	No	Many neurological diseases including epilepsies
PRNP	1	NA	-	AD	No	Prion diseases
SCN1A	1	NA	-	*De novo*	No	Severe myoclonic epilepsy of infancy

PME, progressive myoclonus epilepsy; AR, autosomal recessive; NA, not applicable; AD, autosomal dominant
[1]Applies to individuals with variants in genes previously linked to PMEs.
[2]Variants were classified as probably pathogenic

5.1.3 Analysis of heterozygous variants identified a recurrent *de novo* mutation c.959G>A (p.Arg320His) in *KCNC1* as a worldwide cause of progressive myoclonus epilepsy

To identify new PME genes among cases without pathogenic variants in known disease genes (n=69), we analysed the exome variant data for rare potentially deleterious recessive variants and novel heterozygous variants (*de novo* or autosomal dominantly inherited) in genes previously not established in human disease.

Since the majority of cases were sporadic with only two unsolved cases from 'dominant' families after the analysis of known genes in the cohort, we hypothesised that any underlying pathogenic heterozygous variants would

mostly be *de novo*. However, since parents were not exome sequenced, we could not assess *de novo* mutations directly from the variant data. As an alternative approach, we included potentially deleterious heterozygous variants absent from population variant databases. We ranked the genes based on the number of PME cases with qualifying variants in that gene.

The highest number of cases with variants in a single gene passing the filtering was 11. This number was reached for *KCNC1* and *TTN*, but the latter gene was ruled out from follow-up given that it is highly polymorphic (longest coding region of all human genes) and associated to muscle disorders. We focused on *KCNC1*, where, remarkably, all cases were heterozygous for the same missense variant c.959G>A (variant nomenclature based on Ensembl transcript ENST00000265969.6), corresponding to p.Arg320His on the protein level. *KCNC1* encodes potassium voltage-gated channel (K_V) subfamily C member 1 (official symbol KCNC1, also known as $K_V3.1$ which is used here).

We analysed segregation of the c.959G>A variant in eight families with DNA available and the variant occurred *de novo* in all (I, **Figure 2a**), strongly suggesting it is pathogenic. Interestingly, one index case positive for the variant in *KCNC1* had an affected sibling and two affected children who all had the variant (I, **Figure 2a**). Since the parents of the index case were healthy and negative for the variant, it suggests that one of the parents is a germline mosaic.

Additionally, we screened a secondary cohort of 28 individuals for the *KCNC1* variant. Two additional cases were identified and in one of them, we were able to analyse the parents; the variant occurred *de novo* also in this patient. In total, we identified 16 patients from 13 unrelated families with the variant.

After the initial study (I), we have identified one Italian parent-offspring pair with the c.959G>A variant in *KCNC1*, becoming thus the second familial case (K. Oliver *et al.* unpublished data). In addition, we have become aware of three individuals genotyped in other laboratories in Poland, Italy and USA (**Figure 11**).

To estimate the mutation rate of c.959G>A in *KCNC1*, we used a recently developed method that takes into account the local sequence context at the mutation site when estimating the mutation rate of any given single-nucleotide substitution (Samocha *et al.* 2014). The rate was estimated to be 1.75×10^{-7} mutations per person, which equals to one mutation in every 5,700,000 conceptions. The estimated rate is typical for mutations occurring at CpG dinucleotide sites, the mutation hotspots of the genome, while the average *de novo* mutation rate is lower, 1.20×10^{-8} per nucleotide per generation (Kong *et al.* 2012) (see below section 5.1.3.3 for discussion of the mutation mechanism). This estimate suggests that the mutation potentially affects hundreds of individuals globally. Our observations of the number of *KCNC1* mutation positive individuals in Italy, where a multicenter collaboration between epilepsy clinics has been established (Franceschetti *et*

al. 2014), support this estimate. As of February 2016 we have identified six Italian cases, while based on the mutation rate estimate and the population size of Italians the total number of patients in Italy should be ~10.

Figure 11 The world map of patients with the recurrent c.959G>A mutation in *KCNC1*. Red pins indicate unrelated patients identified in study I (n=13), blue pins those diagnosed after the publication (n=4). Map modified from http://www.vectorworldmap.com/.

5.1.3.1 Arg320His has a dominant-negative effect on K_V3.1 channel function

The p.Arg320His substitution occurs in a highly conserved voltage-sensing segment of the K_V3.1 potassium channel (I, **Figure 2b,c**). The four *in silico* methods used predicted it to be deleterious.

To study the consequence of the variant to K_V3.1 channel function, we collaborated with H. Lerche's research group in University of Tübingen, Germany. The mutated K_V3.1 was analysed in *Xenopus laevis* (African clawed frog) oocytes, of which large size makes it a convenient system to injection of recombinant RNA and to measure ion channel currents by a patch clamp technique (Tammaro *et al.* 2009). Another factor explaining the wide use of *Xenopus* oocytes in basic studies of ion channel function is the lack of significant contamination from endogenous expression of ion channels.

Analysis of Arg320His K_V3.1 channel revealed that it does not produce almost any potassium currents when expressed alone in *Xenopus* oocytes. On the contrary, the wild-type channel generated strong currents upon membrane repolarisation (I, **Figure 3a,b**). To mimic the conditions in patient cells, next, the coexpression of mutant and wild-type channels in 1:1 ratio was examined. An approximately of 80% reduction in potassium currents was observed, indicating a dominant-negative effect for the mutant (I, **Figure 3d,e**). K_V3.1 and other voltage-gated potassium channels function as tetramers consisting of four subunits (see the next section 5.1.3.2 of more detailed coverage of K_V3.1 function). Dominant-negative mutations in these genes have the capability of suppressing the function of the wild-type subunits (Hübner and Jentsch 2002). Finally, the experiments in *Xenopus*

oocytes suggested that the reduced currents of the mutant $K_V3.1$ channel (coexpression of mutant and wild-type) exhibit altered gating properties by showing a hyperpolarising shift in the activation curve (opens at more negative voltages) (I, **Figure 3f**). However, the practical importance of this alteration is likely to be low in relation to the large reduction in potassium currents.

In this study, we did not test whether trafficking of the mutant $K_V3.1$ channels to cell surface is affected, as has been showed for other mutant K_V channels (e.g., Zhao *et al.* 2013). This should, however, be tested in cells of mammalian origin, since trafficking of ion channel subunits occurs in a different fashion in *Xenopus* oocytes, which could affect the results (Tammaro *et al.* 2009).

5.1.3.2 Dysfunction of the $K_V3.1$ channel may cause decreased neuronal inhibition

KCNC1 encodes $K_V3.1$, which functions as a highly evolutionarily conserved K^+ channel subunit that belongs to the K_V3 subfamily of tetrameric voltage-gated K^+ channels ($K_V3.1$–$K_V3.4$, encoded by genes *KCNC1–KCNC4*) (Ried *et al.* 1993). $K_V3.1$ has two isoforms, Kv3.1a and Kv3.1b, which differ in the cytoplasmic C-terminal domain. The isoforms have identical biophysical properties but different subcellular localisation (Gu *et al.* 2012). Variants in *KCNC1* have not been associated with human disease until now, but autosomal dominantly inherited or *de novo* missense variants affecting the highly homologous $K_V3.3$ subunit cause spinocerebellar ataxia 13 (Waters *et al.* 2006; Figueroa *et al.* 2010; Figueroa *et al.* 2011; Németh *et al.* 2013). Numerous other voltage-gated potassium channels are also linked to human disease affecting CNS and other tissues, highlighting their importance for many biological functions (reviewed by Tian *et al.* (2014)).

K_V3 channel subunits can assemble heterotetramers in heterologous expression systems, and this may also occur in brain as subunit expression patterns overlap (Rudy and McBain 2001). K_V subunits consist of six membrane-spanning segments (S1–S6). S4 forms the main voltage sensor where positively-charged residues, in particular four arginine residues occurring every third position, contribute to the gating charge (Aggarwal and MacKinnon 1996; Seoh *et al.* 1996). The PME-causing p.Arg320His substitution in $K_V3.1$ affects the last of these four arginine residues (I, **Figure 2b,c**). Highlighting the functional significance of the voltage-sensor, numerous pathogenic variants in the S4 segment of potassium, sodium and calcium ion channels, have been described in various neurological and muscle disorders (Cannon 2010; Delemotte *et al.* 2010).

Since K_V3 subunits are able to assemble with each other, it is likely that the dominant-negative p.Arg320His in $K_V3.1$ disrupts all K_V3-mediated currents where it is expressed. Studies on mouse models of the K_V3 family members suggest that they are functionally redundant. *Kcnc1* and *Kcnc3* knockout mice present with relatively mild phenotypes, whereas double

mutant mice show prominent symptoms, including myoclonus, tremor, and gait ataxia (Ho *et al.* 1997; Joho *et al.* 1999; Espinosa *et al.* 2001). Thus, the effect of the dominant-negative $K_V3.1$ and $K_V3.3$ mutations seems to be comparable to that of the double knockout. Despite similar functions of $K_V3.1$ and $K_V3.3$, the associated phenotypes are different: myoclonic seizures dominate in $K_V3.1$ dysfunction caused by p.Arg320His and ataxia is the primary phenotype of the mutated $K_V3.3$ channel, even though *KCNC3* positive patients with seizures have also been reported (Waters *et al.* 2006; Figueroa *et al.* 2010; Figueroa *et al.* 2011). Differences in the symptoms likely reflect the distinct expression patterns and biophysical characteristics of the channels (Rudy and McBain 2001).

Voltage-gated potassium channels are transmembrane proteins that have a critical function in transmitting electrical signals, i.e., action potentials, along CNS neurons. They open and close as response to changes in transmembrane potential and return the depolarized cell to the resting state by allowing K+ ions move out of the cell. K_V3 channels have a specific function as the determinants of high-frequency firing of action potentials in several types of CNS neurons, owing to their positively shifted voltage-dependent activation and capability to activate and deactivate at higher rate than other K_V channels (Rudy and McBain 2001). $K_V3.1$ protein is primarily expressed in the CNS, a subpopulation of T lymphocytes being an exception (Gan and Kaczmarek 1998). It is expressed in specific fast-spiking neuron populations in several brain regions. Within the neurons, it can be localised in somata, axons and terminals (Sekirnjak *et al.* 1997; Ozaita *et al.* 2002). Notably $K_V3.1$ has high expression in inhibitory GABAergic interneurons (Gan and Kaczmarek 1998; Rudy and McBain 2001). Interneurons are typically local-circuit neurons that pass information between other neurons (Markram *et al.* 2004). They are mostly inhibitory because they generally use GABA as a neurotransmitter. Given the important role of $K_V3.1$ in cortical interneurons, we hypothesise that decreased inhibition of neuronal activity would result from the p.Arg320His substitution in $K_V3.1$ due to impaired firing of fast-spiking inhibitory GABAergic interneurons (**Figure 12**). This mechanism is likely to increase neuronal excitation and predispose to myoclonus and tonic-clonic seizures seen in the patients. As mentioned in section 2.4.3, failure of inhibitory control in the CNS has been suggested to contribute to other forms of epilepsies as well (Yu *et al.* 2006).

Supporting our hypothesis of decreased inhibition as the underlying epileptogenic mechanism of K_V3 channel dysfunction, increased cortical excitability and susceptibility to seizures in *Kcnc2* knockout mouse has been proposed (Lau *et al.* 2000) to result from the suppression of inhibitory interneurons. *Kcnc1* knockout mice do not show increased seizure activity (Ho *et al.* 1997) but exhibit impaired motor skill (Ho *et al.* 1997) and altered oscillations recorded at the somatomotor cortex (Joho *et al.* 1999). Furthermore, impaired function of cerebellar neurons, where $K_V3.1$ is expressed (Gan and Kaczmarek 1998) is likely to contribute to the motor

impairment. *In vivo* model systems are required to assess the consequence of p.Arg320His on neuronal and network physiology and motor function.

Activation of K_V3 channel function would be a potential therapeutic approach in patients with *KCNC1* mutations. Interestingly, a UK-based biotechnology company Autifony Therapeutics Ltd. has developed compounds acting as positive modulators of the $K_V3.1$ channels (Alvaro and Marasco 2013; Rosato-Siri *et al.* 2015; Taskin *et al.* 2015). One of these molecules is currently on phase IIa clinical trial in the treatment of tinnitus and one on phase I trial in the treatment of schizophrenia. We are currently collaborating with the company to assess the effect of these compounds on the mutant $K_V3.1$ channel.

Importantly, $K_V3.1$ was recently established as the predominant voltage-gated K^+ channel in adult neural precursor cells, involved in neurogenesis of adult brain (Yasuda *et al.* 2013). *KCNC1* gene knockdown with specific small interfering RNAs in adult neural precursor cells decreased neural precursor cell proliferation and neuronal differentiation. Hence, the authors suggested that selective $K_V3.1$ activation could be a potential therapeutic approach for neurodegenerative diseases. Our observation of $K_V3.1$ dysfunction causing a neurodegenerative disease provides a direct target to test this hypothesis.

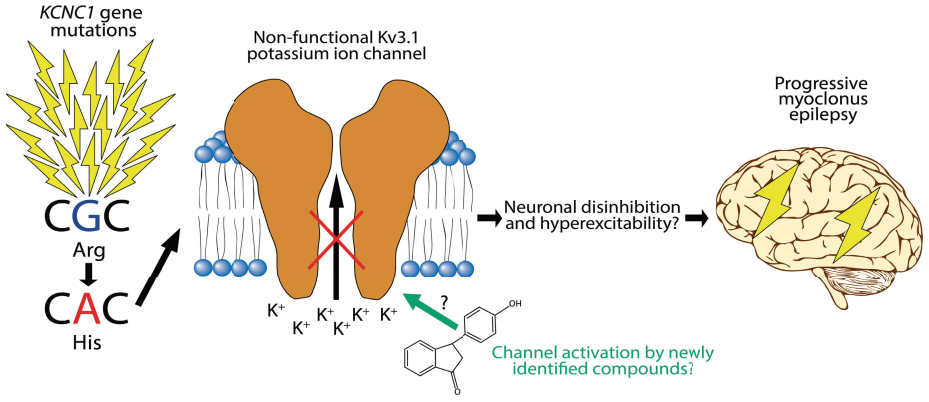


Figure 12 A cartoon of the underlying mechanisms of progressive myoclonus epilepsy caused by a *de novo* mutation *KCNC1*. Mutations (lightning bolts) hit the same nucleotide in *KCNC1* and disrupt the function of a brain-expressed potassium ion channel, which causes a severe form of epilepsy due to decreased inhibition. A potential therapeutic approach is to use newly identified compounds to activate the $K_V3.1$ potassium channel.

5.1.3.3 Why the c.959G>A mutation in KCNC1 is recurrent and why we have not observed any other changes in KCNC1?

As mentioned above, the identification of *de novo* mutations as a cause of approximately half of the solved PME cases in the exome sequencing cohort was unexpected, given that the vast majority of previously established causes of PME have shown autosomal recessive inheritance (Franceschetti *et al.*

2014). An even more surprising observation was that a single *de novo* mutation in *KCNC1* underlies 13.0% of the exome cohort. Notably, no other pathogenic changes have been observed to date in the gene, while four additional cases with the recurrent c.959G>A have been identified after publication of study I.

Observing recurrent mutations in epilepsies or other neurological diseases is not a rare phenomenon (Ducros *et al.* 1999; Kearney *et al.* 2006; Wagnon and Meisler 2015) but reports of diseases where the vast majority cases are caused by a single variant occur substantially less often (e.g., Simons *et al.* 2013). This phenomenon of a single, highly recurrent mutation explaining the majority of or all cases has been reported in association with distinguishable syndromes including achondroplasia (mutated gene *FGFR3*) (Rousseau *et al.* 1994; Shiang *et al.* 1994), Muenke syndrome (*FGFR3*) (Bellus *et al.* 1996), Apert syndrome (*FGFR2*) (Wilkie *et al.* 1995), and Hutchinson-Gilford progeria syndrome (*LMNA*) (Eriksson *et al.* 2003).

At least two factors may account for the recurrent occurrence of c.959G>A mutation in *KCNC1*. First, the mutation occurs in a CpG dinucleotide, which are known mutation hotspots (Kong *et al.* 2012) due to the spontaneous deamination of methylated cytosines to thymine (Pfeifer 2006). Notably, the recurrent mutations in the above mentioned syndromes occur in CpG sites.

Second, *KCNC1* mutations other than c.959G>A may cause a phenotype different from PME. This might be the single most important explanation why no other mutations have been observed in our PME cohort. This hypothesis is supported by observations in ataxia caused by mutations in *KCNC3*. In the case of *KCNC3*, missense changes encoding p.Arg420His and p.Arg423His in the S4 segment of the K_V3 channel subunit (I, **Figure 2c**) are associated with different clinical manifestations, late-onset progressive ataxia and early-onset slowly progressive ataxia, respectively (Figueroa *et al.* 2010). The effects of these mutations on $K_V3.3$ channel function are distinct (Minassian *et al.* 2012; Zhao *et al.* 2013), and the amino-acid specific consequences of mutations occurring in the voltage-sensor segment can be generalised to other voltage-gated potassium channels as well (Seoh *et al.* 1996). The more severe consequence of p.Arg423His in *KCNC3*, which is occurs in a position analogous to the PME-causing substitution p.Arg320His in *KCNC1*, is likely due to both a dominant-negative suppression of current amplitude and a dominant gain-of-function effect on channel gating (Minassian *et al.* 2012), similarly than we showed in this study for p.Arg320His in *KCNC1*. Notably, a recurrent occurrence of both the p.Arg420His (Waters *et al.* 2006; Figueroa *et al.* 2010) and p.Arg423His (Figueroa *et al.* 2010; Figueroa *et al.* 2011) mutations in *KCNC3* has also been reported. Indeed, the S4 segments of voltage-gated ion channels contain highly conserved and essential positively charged arginines encoded by CpG containing codons, and therefore it is not surprising that S4 segment arginine mutations underlie also several other channelopathies (Cannon 2010; Delemotte *et al.* 2010).

Finally, one explanation for the lack of other identified pathogenic variants in *KCNC1* could be that other missense variants or loss-of-function changes in the gene are mostly tolerated. However, this is likely not the case, because *KCNC1* is among the top 1% least tolerant genes for functional variation (Samocha *et al.* 2014; Lek *et al.* 2015), suggesting that variants throughout the gene have phenotypic consequences. To conclude, observing c.959G>A as the sole *KCNC1* mutation in our study may be primarily because other mutations in the gene have slightly different biophysical consequences and the resulting phenotype would not meet all diagnostic criteria for PME.

To study the mutation mechanism of the recurrent c.959G>A mutation in *KCNC1*, we are currently assessing the parent of origin of the mutation. Preliminary data from two individuals shows that in one patient the mutation originated from the mother and in one from the father (M. Muona, unpublished data), suggesting that mutations do not occur at least exclusively in males. Notably, mutations in the above-mentioned syndromes caused by single mutations in CpG sites occur exclusively in fathers (Glaser and Morison 2009). In diseases showing a more diverse mutational spectrum, such as Dravet syndrome, majority, but not all, of the *de novo* variants are of paternal origin (Heron et al. 2010). In general, the majority of *de novo* variation arises in males, the number of *de novo* variants increasing along with father's age (Kong et al. 2012). Indeed, spermatogenesis involving a higher number of cell divisions is more susceptible to point mutations compared to oogenesis. Furthermore, the exclusive occurrence of achondroplasia and Apert syndrome mutations in males, is at least partially explained by the observation that spermatogonial stem cells with the recurrent mutations have a proliferative advantage over unmutated cells (Goriely et al. 2003; Shinde et al. 2013).

5.1.3.4 Myoclonus epilepsy and ataxia due to potassium channel mutation (MEAK) resembles clinically the most common form of progressive myoclonus epilepsy

We proposed the newly identified form of PME to be termed as myoclonus epilepsy and ataxia due to potassium channel mutation (MEAK). The main clinical features of the 16 MEAK patients identified in study I are shown in Table 1 of the original publication. Briefly, the first symptoms, which are usually myoclonic seizures, emerged at ages of 5–14 years. Myoclonus typically became debilitating during adolescence. Ataxia is observed in all. Learning disability and cognitive decline was noted in some cases. MRI findings were normal or show cerebellar atrophy. Premature death has not been observed. The oldest known MEAK case, identified after the initial publication, is now 63 years of age. Clinical presentation was somewhat milder in the family where the mutation was transmitted from the affected mother to offspring (I, **Figure 2a**). However, in most cases the phenotype is severe, explaining why the mutation occurs primarily *de novo*.

Interestingly, MEAK has substantial clinical overlap with ULD, the most common form of PME, highlighting the importance of being able to perform a molecular diagnosis in these diseases. The most apparent difference in the clinical presentation between ULD and MEAK is that the clinical course of MEAK is typically more severe, with patients requiring a walking aid by their teens, whereas ULD patients more often remain ambulant with the disorder stabilizing in middle age (Magaudda *et al.* 2006).

Given that the underlying molecular biology for these two forms of PME appears quite different, it is intriguing to see clinical overlap between them. ULD, which is autosomal recessively inherited, is caused by mutations in the gene encoding cystatin B (CSTB), a cysteine protease inhibitor with roles implicated in apoptosis, oxidative stress and inflammation (Pennacchio *et al.* 1998; Lehtinen *et al.* 2009; Tegelberg *et al.* 2012; Okuneva *et al.* 2015). Recent evidence suggests that altered GABAergic signaling contributes to the pathogenesis in *Cstb* knockout mice (Buzzi *et al.* 2012; Joensuu *et al.* 2014), which may underlie hyperexcitability predisposing to myoclonus and seizures. Moreover, loss of GABAergic neurons have been observed in the mouse model for Lafora disease (Ortolano *et al.* 2014). These observations suggest a possible convergent pathway for PMEs associated with variants in different genes.

5.1.4 Exome variant data of unsolved patients contain genes of potential interest for further research

We did not identify any candidates as new recessive disease genes using our criteria of observing variants in a gene in at least two cases under the recessive approach (I, **Figure 1a**). Five genes with qualifying variants in two or more unsolved PME cases were identified but they were excluded as new epilepsy genes since the genes were, for example, highly polymorphic, linked to muscle disorders and/or not expressed in the CNS (I, **Supplementary Table 7**).

To identify potential disease genes for further research, we analysed the variant data for biallelic loss-of-function variants (nonsense, canonical splice site, frameshift indel). We observed nine such genes in single cases (I, **Supplementary Table 8**). Out of these, we considered *ALG10* and *APOA1BP* in particular as genes of potential interest as they encode functionally conserved proteins expressed in the CNS. These genes are among those where we are actively looking for additional cases through collaboration, including the GeneMatcher data sharing platform. Identification of a second, independent hit in another patient, followed by functional studies, would be the next steps of establishing them as new disease genes.

In the analysis to identify novel dominant/*de novo* genes other than *KCNC1*, we observed 14 genes with potential variants in four or more patients who did not have the c.959G>A mutation in *KCNC1* or mutations in known disease genes (I, **Supplementary Table 2**). However, we did not consider

any of the genes as likely true PME disease genes for similar reasons described above for the excluded recessive genes. In addition, we discarded some of the genes, since variants were inherited from unaffected parents in the segregation analysis. Since there were 425 genes with two to three cases with variants passing the filtering, we were not able to pinpoint obvious novel genes in this category.

5.1.5 Ways to dissect still unsolved PME cases

Overall, we solved 31.0% of the study cohort, thus reducing the proportion of molecularly undiagnosed PME cases, which is estimated to be 28% in the Italian population (Franceschetti *et al.* 2014). The success rate is in line with several other WES studies aiming to elucidate the underlying genetic causes in various rare disorders (Iglesias *et al.* 2014; Srivastava *et al.* 2014; Yang *et al.* 2014; Retterer *et al.* 2015; The Deciphering Developmental Disorders Study 2015; Helbig *et al.* 2016; Powis *et al.* 2016).

Two-thirds of the cohort remained without a molecular diagnosis. The unsolved patients were heterogeneous in terms of age of onset and associated clinical features, as were the solved cases. Family history of the unsolved cases is also heterogeneous and includes 50 sporadic cases (50/70, 71%), seven cases from recessive families (7/11, 64%), and one case from dominant families (1/3, 33%) (I, **Figure 1b**).

Given that we now established an important role for *de novo* mutations in PMEs, we are next performing WES in trio setting for unsolved cases. This approach has been successful in many other epilepsy WES studies (e.g., Epi4K Consortium and Epilepsy Phenome/Genome Project 2013; EuroEPINOMICS-RES Consortium *et al.* 2014). This study was limited to coding regions of the human genome, and even there coverage was not complete since ~8% of the exome was captured with less than 10 reads. It is likely that WGS would solve some proportion of the remaining individuals, which has been the case in, for example, intellectual disability (Gilissen *et al.* 2014). We also did not assess CNVs or variants showing di- or polygenic inheritance. Preliminary analysis of the current exome sequence data for CNVs indicates that the data do not contain CNVs in known epilepsy genes, while exome-wide investigation of CNVs remains to be carried out (M. Muona, unpublished data). Finally, it is likely that pathogenic variants in some proportion of the unsolved individuals were already captured by the current WES data. However, the clinical significance of these variants was not appreciated, since they occurred in a new disease gene and in one individual only, or alternatively, they were discarded in the analysis pipeline (e.g., synonymous variants). As mentioned above, data sharing and further genetic studies in PMEs and related disorders may assist pinpointing further causal variants in the existing WES data.

5.2 Biallelic variants in *UBA5* cause an infantile-onset encephalopathy (II)

5.2.1 Analysis of in-house and external exome datasets revealed compound heterozygous variants in *UBA5* in five families with similar clinical features

As part of a WES project aiming to identify the underlying causes of 30 Finnish individuals having severe infantile-onset epileptic syndromes, we analysed family A with two affected siblings (**Figure 10** p. 61; II, **Figure 1B**). In this family, healthy parents and the index case, assigned here as A-4, were sequenced. Given that the affected siblings represent both sexes, we primarily searched for rare, potentially deleterious autosomal recessively inherited variants.

The only variants qualifying the autosomal recessive variant filtering were compound heterozygous variants in *UBA5* (Ubiquitin-like modifier activating enzyme 5). One of the variants is c.1111G>A, which encodes a missense substitution p.Ala371Thr (variant nomenclature based on GenBank ID NM_024818.3). The p.Ala371Thr is relatively common in the ExAC data: carrier frequency is 0.92% in Finns and 0.56% in other Europeans, with no homozygotes, however, detected. The other variant identified in family A, c.164G>A, was predicted to encode an amino acid substitution p.Arg55His, but later analysis of *UBA5* in the RNA level in patient cells showed that this change also affects splicing (see section 5.2.2). This variant is very rare, seen only once in the reference databases with one carrier in ExAC. Both variants occur in evolutionarily conserved sites (II, **Figure 1E**) and are predicted to be deleterious by the used *in silico* methods. Segregation analysis by Sanger sequencing showed that the variants segregate in an autosomal recessive manner in this family with two unaffected siblings (II, **Figure 1B**).

In addition to the 30 patients sequenced at WTSI, we have exome sequenced and analysed in-house six additional families with severe epileptic syndromes and ascertained through collaboration. One of these is a family ascertained by clinicians in University of Oulu, Finland (denoted as family B). The family has four affected and six unaffected siblings born to healthy parents. Exome sequencing was performed for the parents, two affected siblings and one unaffected sibling (**Figure 10**, p. 61; II, **Figure 1B**). Also in this family, the only variants passing the filtering criteria of autosomal recessive analysis were compound heterozygous variants in *UBA5*. Of the two variants identified in this family, one was the same as an in Family A, p.Ala371Thr. The other was a nonsense variant c.855C>A (p.Tyr285Ter), occurring in exon 9 of the 11 exons of *UBA5*, and predicted to cause nonsense mediated decay making the allele dysfunctional. Sanger sequencing of the variants in one additional affected sibling and two unaffected siblings confirmed the autosomal recessive segregation of these variants in family B (II, **Figure 1B**).

Next, we utilised externally generated datasets to try to identify additional families with biallelic variants in *UBA5*. We accessed GeneMatcher web resource as well as contacted researchers at the EuroEPINOMICS Rare Epilepy Syndrome project (178 WES or WGS samples), Deciphering Developmental Disorders (DDD) study (exomes from >4,000 families with developmental disorders; The Deciphering Developmental Disorders Study 2015) and Northern Finland Intellectual Disability cohort (455 independent exomes). In the DDD and Northern Finland Intellectual Disability exomes, we identified additional three unrelated individuals who are clinically similar to patients in families A and B and have compound heterozygous variants in *UBA5*. These three patients (C-4, D-3 and E-3) have p.Ala371Thr *in trans* with a novel or very rare nonsense variant (II, **Figure 1B**).

Given that the p.Ala371Thr variant, which is present in all five families, is relatively common in Europeans, we assessed the probability to observe rare, biallelic *UBA5* variants in multiple families only by chance in the analysed study populations. We utilised a recently established statistical tool, 'recessiveStats' (Akawi *et al.* 2015), which can be used to pinpoint genes that are enriched for biallelic functional (e.g. missense, in-frame indel) and loss-of-function variants in the study cohorts. For each protein-coding gene, the tool fetches the ExAC database to determine cumulative allele frequencies of functional and loss-of-funcion variants in the general population. By using the obtained frequencies recessiveStats evaluates the expected prevalence of biallelic genotypes in a given gene, which is then compared to that observed in the study. Using this tool we showed that it is highly unlikely to observe only by chance four (B-E) families with missense + loss-of-function *UBA5* variants (variants of family A are annotated as missense+missense) in our study populations (P=3.30×10^{-10}). Moreover, we accessed over 75,000 control exomes in the ExAC, SISu and DDD studies and observed no individuals with p.Ala371Thr *in trans* with a loss-of-function variant, further supporting a pathogenic role for this allele combination.

Clinical evaluation of the nine patients in five families showed substantial overlap. Briefly, the symptoms emerged within the first months of life with irritability, which was followed in most by dystonic movements, epileptic seizures and truncal hypotonia. The patients showed moderate to severe intellectual disability and stagnation of development. Patients in family B also had severe growth retardation. Patients showed microcephaly but MRI and neuropathological findings were mild and unspecific. Premature death is associated with the disease. Clinical features are described in more detail in the manuscript of study II.

Segregation and statistical data of these rare *UBA5* variants, and the substantial clinical overlap between the nine affected individuals in five unrelated families together strongly suggest that these variants are the cause of the newly identified syndrome. Further supporting this conclusion, we became aware of another research group (D. Bonneau *et al.*, personal

communication) who has characterized five patients in four families with compound heterozygous variants in *UBA5* and similar clinical presentation.

5.2.2 Functional studies show that identified *UBA5* variants have a hypomorphic effect on the enzymatic activity of UBA5

We performed several experiments to elucidate the molecular consequences of the identified variants. *UBA5* encodes an activating enzyme for UFM1, which is a ubiquitin-like protein (UBL) that is ligated to its target proteins with a covalent bond (Komatsu *et al.* 2004). UFM1 and other UBLs act as post-translational regulators that have functions in various cellular processes (more information of the function of the UFM1-system is presented in the next section 5.2.3). The first step of the activation of UFM1 by UBA5 is the formation of a noncovalent complex between UBA5 and UFM1 (Komatsu *et al.* 2004; Bacik *et al.* 2010) (II, **Figure 1A**). Then a thioester bond between the catalytically active cysteine residue 250 of UBA5 and a C-terminal glycine residue in UFM1 is formed. After UFM1 activation, UBA5 interacts with the conjugating enzyme of UFM1, UFC1, and transfers the activated UFM1 from UBA5 to UFC1 (Komatsu *et al.* 2004). Finally, UFL1 enzyme ligates UFM1 to its target protein (Tatsumi *et al.* 2010).

First, we studied the identified *UBA5* variants on the RNA level. Analysis of *UBA5* cDNA in fibroblasts of one of the affected individuals in family A, showed that the c.164G>A variant predicted to cause p.Arg55His substitution affects in fact splicing by facilitating non-frame skipping of exon 2. Since skipping of exon 2 leads to a premature stop codon, the mRNA is likely subjected to nonsense mediated decay, which explains why *UBA5* expression is approximately 30% lower in the patient fibroblasts (II, **Figure 2A**). However, this effect on splicing does not appear to occur with full efficiency leaving some transcripts normally spliced. This means that Arg55His mutant is expressed in patient cells at lower levels, but as shown below, it is nonfunctional. In fibroblasts of one patient from family B, *UBA5* expression is 50% lower due to the fact that the other variant in this family encodes a premature stop codon (II, **Figure 2A**). Indeed, analysis of cDNA in this patient shows that Ala371Thr is the only allele expressed.

Our collaborator Prof. Masaaki Komatsu's lab has established biochemical assays to assess the above described function of UBA5 as the activating enzyme of UFM1. First, UBA5 function was tested in an *in vitro* thioester formation assay, where reactions contained recombinant proteins UBA5 (wild-type or one of the two missense mutants Arg55His or Ala371Thr), UFM1 and UFC1. This assay showed that both UBA5 mutants have lower ability to form an intermediate with UFM1 compared to the wild-type UBA5 (II, **Figure 3C,D**). In addition, neither UBA5 mutant was able to transfer UFM1 to UFC1 in a short 5-minute assay. When the reaction time was extended up to 60 minutes, Arg55His showed approximately 50% reduction in the ability to transfer UFM1 to UFC1, while for Ala371Thr the reduction was 20-30% (II, **Supplementary Figure S7**). The ability of UBA5 mutants

to activate UFM1 and transfer it to UFC1 was tested also in HEK293T cells depleted with endogenous *UBA5*. In these cells, either wild-type or mutant UBA5 was overexpressed together with UFM1 and UFC1. Arg55His mutant showed reduced ability to activate UFM1, and also formation of the UFM1-UFC1 complex was suppressed (II, **Figure 3A,B**). For Ala371Thr mutant, the ability to activate UFM1 was not significantly reduced compared to wild-type, but the ability to transfer UFM1 to UFC1 was weaker in comparison to that of the wild-type. UBA5 function was also examined in primary skin fibroblasts derived from two patients, one from each family. Compared to control fibroblasts, there were less UFM1-UBA5 and UFM1-UFC1 complexes in patient-derived fibroblasts (II, **Figure 2B**). Finally, the effect of UBA5 mutants on UFM1-target protein conjugate formation was examined. With Arg55His both quantity and number of conjugates was clearly reduced, while for Ala371Thr the decrease was only suggestive (II, **Figure 3E,F**)

As a conclusion, the biochemical experiments suggest that Arg55His mutant has severely attenuated ability to activate UFM1 and transfer it to UFC1, whereas in Ala371Thr mutant, the defect is milder. Compatible with the hypomorphic nature of the Ala371Thr mutant, UBA5-UFM1 and UFM1-UFC1 intermediates are not completely absent from patient fibroblasts, indicating there is some UBA5 activity remaining.

Based on the functional experiments, the nine patients in five families are heterozygotes for a variant of mild effect (p.Ala371Thr) and a variant of a severe, loss-of-function effect on UBA5 activity (combined splice and severe missense variant in family A and nonsense variants in families B, C, D and E). In other words, on the protein level the only residual UBA5 enzymatic activity comes from the p.Ala371Thr allele in all families. This likely explains the substantial clinical overlap between the families. Noteworthy, the other research group with *UBA5* patients has identified two families of European origin who are genotypically similar to ours, since they have the p.Ala371Thr variant combined with a loss-of-function variant (D. Bonneau *et al.*, personal communication).

5.2.3 Pathomechanism of UBA5 dysfunction

Ubiquitin and UBLs are small proteins that function as post-translational regulators of various cellular processes. UFM1 is a highly conserved UBL that can be found in all multicellular organisms. Members of the UFM1 cascade are ubiquitously expressed. The function of the UFM1-conjugation is, however, largely unknown and only two target proteins, UFBP1 and ASC1, have been identified (Tatsumi *et al.* 2010; Yoo *et al.* 2014).

Highlighting their important role in a wide range of cellular tasks, genetic defects and/or abnormal function of members of ubiquitin and UBL pathways have been linked to various genetic disorders, including neurodegenerative and neurodevelopmental disorders, anaemia, muscular atrophy and growth retardation syndrome (Kishino *et al.* 1997; Huber *et al.* 2005; Jacquemont and Taniguchi 2007; Ramser *et al.* 2008; Atkin and

Paulson 2014). Before this study, disease-causing variants in genes encoding proteins of the UFM1-system had not been reported to date, with the exception of a missense variant in *UFSP2* (UFM1-specific peptidase) that shows suggestive association with hip dysplasia (Watson *et al.* 2015). Moreover, while the manuscript of this study was under review, Duan and colleagues reported one family with two siblings with childhood-onset ataxia and compound heterozygous *UBA5* variants (p.Lys310Glu and p.Arg246Ter) (Duan *et al.* 2016). This suggests that *UBA5* variants may also be linked to a substantially milder neurological disease in addition to the infantile-onset encephalopathy described in our study. The severity of *UBA5* associated conditions most likely correlates with the degree of remaining enzymatic activity of UBA5.

Prior to genetic studies in human, the importance of the UFM1-system has been noted by studying mice with knocked out *Uba5*, *Ufm1* or *Ddrgk1* (*Ufbp1*). All of these gene knockouts cause embryonic lethality, showing that UFM1-system function is indispensable for normal development (Tatsumi *et al.* 2011; Cai *et al.* 2015). Furthermore, studies have shown that the UFM1-system is crucial for erythroid development, breast cancer development and endoplasmic reticulum (ER) stress response (Lemaire *et al.* 2011; Tatsumi *et al.* 2011; Zhang *et al.* 2012; Yoo *et al.* 2014). However, the function of the UFM1 system has not been assessed in the CNS.

In this study, we were able to analyse the function of the UFM1-system in the CNS by studying mice with a CNS-specific knockout of *Ufm1* that was generated using Cre-Lox recombination (II, **Figure 4**). The mice died within one day after birth. Neuropathological examinations of the brain of the *Ufm1* mice did not show major irregularities in cellular organization but the brains were microcephalic. Interestingly, neurons in the occipital region of the brain showed increased apoptosis. Findings in mice lacking *Ufm1* in the CNS were similar to those observed in *UBA5* patients, who also had microcephaly but major abnormalities in brain structure were absent. In conclusion, results from the mice lacking *Ufm1* in the CNS show that UFM1-system has an important role for normal CNS development and cell survival.

Observation of apoptosis in the *Ufm1* mouse neurons is intriguing, since silencing of the UFM1-system has been shown to increase ER stress induced cell death in pancreatic beta cells and hematopoietic stem cells (Lemaire *et al.* 2011; Cai *et al.* 2015). In normal conditions, ER takes care of protein folding and the properly folded proteins are transported to the Golgi apparatus (Ozcan and Tabas 2012). However, certain stimuli, such as hypoxia, oxidative injury or some chemicals, may induce ER stress, which is a condition where the normal ER homeostasis is altered, and unfolded proteins accumulate in the ER. Normally, cell copes with this stress by increasing folding capacity of proteins and inhibiting protein translation. This is called the ER stress response or unfolded protein response. However, prolonged or excessive ER stress can trigger cell death. Also, genetic defects disrupting components of the ER stress response pathway may induce cell

death and be involved in various human diseases, including neurodegenerative diseases (Zhao *et al.* 2005; Ozcan and Tabas 2012).

Given the suggested role of the UFM1-system to protect cells from ER stress mediated cell death, it is therefore tempting to hypothesise that in our patients reduced UBA5 activity causes diminished conjugation of UFM1 to its target proteins, which impairs ER stress response and leads to excessive cell death, in the CNS in particular (**Figure 13**). Further studies are warranted to test this hypothesis about the pathomechanism of the *UBA5* associated syndrome. Already now, the other group with *UBA5* families has observed that ER stress response is indeed altered in fibroblasts of patients with biallelic pathogenic variants in *UBA5* (D. Bonneau *et al.*, personal communication). Finally, a knock-in mouse expressing the mouse analogue for the human p.Ala371Thr variant of *UBA5* would be of high value in dissecting the molecular consequences of the attenuated UFM1-system.

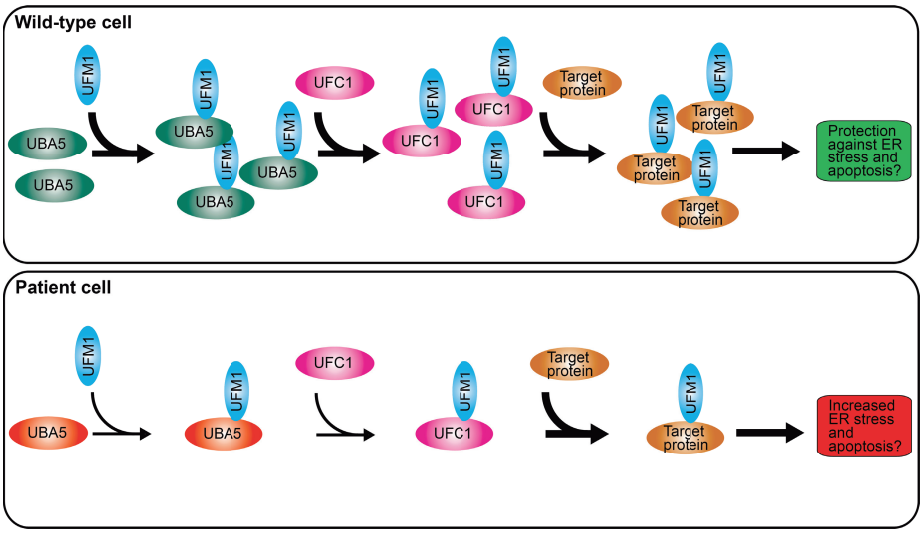


Figure 13 A schematic of the hypothetical pathomechanism of UBA5 dysfunction. In wild-type cells, UBA5 activates UFM1 robustly and transfers UFM1 to UFC1. Upon endoplasmic reticulum (ER) stress induction, this pathway protects cell from ER stress induced apoptosis. In patient cells, UBA5 population is reduced in both quantity and catalytic activity, and consequently, UFM1 activation and transfer of UFM1 to UFC1 occurs less efficiently. Consequently, the number of UFM1-target protein conjugates is reduced, which predisposes to ER stress mediated apoptosis. The figure is modified from Figure 1A (study II), which was created by Masaaki Komatsu.

5.3 Loss-of-function of ADAM22 implicated in rapidly progressing encephalopathy with epilepsy and cortical atrophy (III)

5.3.1 Analysis of whole-exome sequencing data revealed compound heterozygous variants in *ADAM22*

Like family A in study II, an affected individual with rapidly progressing encephalopathy with epilepsy and cortical atrophy was sequenced as part of the WES project consisting of 30 Finnish individuals with severe, infantile-onset epileptic syndromes (see original publication III for more detailed description of clinical features). Analysis of the WES data of the female patient did not reveal any candidate variants in known disease genes or in mtDNA. Similarly, analysis of genome-wide SNP data for CNVs in the parent-offspring trio did not show any likely pathogenic changes.

Next, we looked for novel causes by analysing the exome data for recessive and *de novo* variants. However, parents of the patient were not exome sequenced so were not able to assess *de novo* variants directly. Filtering the variant data for novel heterozygous variants yielded seven variants, which is a manageable number (see below for discussion about the ability to dramatically narrow down the number of variants in exome data from samples of Finnish origin). Segregation of these variants was analysed by Sanger sequencing and all of them turned out to be inherited from one of the parents. Analysis of exome data for recessively inherited variants yielded two genes, *ADAM22* and *CPA4*, but only the compound heterozygous variants in *ADAM22* segregated in an autosomal recessive manner in the subsequent segregation analysis done by Sanger sequencing (III, **Figure 2A**).

Of the compound heterozygous variants in *ADAM22* (ADAM metallopeptidase domain 22), one is a missense variant c.1202G>A (p.Cys401Tyr) that is present in only two carriers in the ExAC database (mutation nomenclature based on GenBank sequence NM_021723). It is predicted to be deleterious by all four prediction programs used. The p.Cys401Tyr variant occurs in a highly conserved cysteine residue that forms a disulfide bond with another cysteine residue at position 394 (III, **Figure 2B,C**) (Liu *et al.* 2009). The amino acid substitution could thus potentially alter the tertiary structure of the protein. The other variant in *ADAM22* is a 1bp frameshift deletion in exon 27 of *ADAM22* with a total of 31 exons (c.2396delG, p.Ser799IlefsTer96). It is predicted to alter the last ~100 amino acids of the full-length 906-amino acid ADAM22 protein (III, **Figure 2B**).

Both filtering approaches, "*de novo"* and recessive, yielded a manageable number of candidate variants, even though we did not exome sequence parents. This applies also to other Finnish exomes that we have generated in-house. In those, we observe on average ~15 heterozygous variants per exome that pass the filtering for novel, potentially deleterious heterozygous variants. On the contrary, in non-Finnish Europeans using the same filtering criteria

as in this study, we see a substantially higher number (~50 on average). This difference indicates that variant filtering in the Finnish exomes benefits a lot from the comprehensive representation of Finns in the variant databases (3,200 Finns in ExAC and 100 in the 1000 genomes) and from the reduced genetic heterogeneity in the Finnish population due to past bottlenecks in the population history (Peltonen *et al.* 1999).

We aimed to identify additional individuals with severe epileptic syndromes and recessive *ADAM22* variants in various WES cohorts. We did not identify any additional cases in the in-house exomes, GeneMatcher database, 178 epileptic encephalopathy exomes of the EuroEPINOMICS project or in first ~1000 trios of the Deciphering Developmental Disorders project.

5.3.2 Identified ADAM22 mutants do not bind to its ligand, LGI1

Although we did not identify a 'second hit', i.e., another case with pathogenic *ADAM22* variants, which is the prerequisite for the establishment of a new disease gene, we progressed with functional studies of the identified *ADAM22* variants. This was done because ADAM22 has been under active research among epilepsy researchers and neurobiologists for the past decade as it has been recognized as the postsynaptic receptor for a secreted glycoprotein LGI1 (Fukata *et al.* 2006), where mutations cause autosomal dominant lateral temporal lobe epilepsy (Kalachikov *et al.* 2002). Moreover, since *Adam22*$^{-/-}$ mice present with lethal seizures (Sagane *et al.* 2005), we considered that the identified *ADAM22* variants are a good candidate for the genetic cause of the disease in our patient.

Our collaborators in Prof. Masaki Fukata's lab studied whether the two ADAM22 mutants can bind to its two known binding partners, LGI1 and PSD-95, of which the latter is a postsynaptic scaffolding protein. Wild-type ADAM22 binds to LGI1 at the postsynaptic membrane. Interaction with PSD-95 occurs in the cytoplasm where the C-terminal PDZ-binding motif of ADAM22 binds to the PDZ-domain of PSD-95 (Lovero *et al.* 2015). A cell-surface binding assay indicated that neither the Cys401Tyr missense mutant nor the frameshift mutant is able to bind LGI1 in a heterologous expression system in COS7 cells (III, **Figure 3A**). Furthermore, immunoprecipitation assay in HEK293T suggested that neither mutant coimmunoprecipitates with LGI1 (III, **Figure 3C**). The frameshift mutant did not also bind to PSD-95, since the mutant misses the C-terminal PDZ-binding motif (III, **Figure 3D**). Taken together, these experiments suggest that both ADAM22 mutants are dysfunctional and cannot bind to LGI1, and in the case of the frameshift mutant, ability to bind to PSD-95 is also diminished.

5.3.3 Pathomechanism of ADAM22 dysfunction

ADAM22 is expressed at the postsynaptic membrane of certain CNS neurons (Fukata *et al.* 2006; Fukata *et al.* 2010; Lovero *et al.* 2015) (**Figure 14**). It is

also expressed in axons and peripheral nervous system (Ogawa *et al.* 2010; Özkaynak *et al.* 2010), but its role in the postsynaptic membrane has been studied most extensively. The LGI1-ADAM22 complex regulates development and functional maturation of synapses in early postnatal stages and plays a role in maintaining the brain excitability throughout the life (Zhou *et al.* 2009; Boillot *et al.* 2014; Lovero *et al.* 2015). Specifically, the LGI1-ADAM22 complex recruits AMPA-type glutamate receptors to the postsynapses, which is one of the hallmarks of excitatory synapse maturation (Fukata *et al.* 2006; Fukata *et al.* 2010; Lovero *et al.* 2015). The interaction between the C-terminal PDZ-binding motif of ADAM22 and the PDZ-domains of PSD-95 is required for the regulation of the AMPA receptors (Lovero *et al.* 2015). Another hallmark of glutamatergic synapse maturation is the change of subunit composition of the NMDA-type glutamate receptor. It has been suggested that when LGI1 is mutated, the ratio of NR2B/NR2A subunits of the NMDA-receptor remain high, which increases the excitatory effect of glutamate (Zhou *et al.* 2009). As with AMPA receptors, the interaction between ADAM22 and PSD-95 regulates NMDA receptor function (Lovero *et al.* 2015). Further stressing the role of LGI1 in downregulation of excitation, targeted deletion of LGI1 in glutamatergic excitatory neurons has been shown to be sufficient to generate seizures (Boillot *et al.* 2014). Finally, highlighting the importance of LGI1-ADAM22 complex in keeping brain excitability in balance throughout life, one form of autoimmune limbic encephalitis, which involves amnesia and seizures, is due to autoantibodies against LGI1 (Irani *et al.* 2010; Lai *et al.* 2010). This has been shown to cause disruption of LGI1-ADAM22 interaction and subsequent reduction of AMPA receptors (Ohkawa *et al.* 2013). Given that the LGI1-ADAM22 complex is also expressed in inhibitory interneurons, whose excitatory input is mediated by AMPA-type glutamate receptors, the authors suggested that reduced inhibition of neural networks due to lowered excitation of interneurons is one potential epileptogenic mechanisms related to LGI1-ADAM22 dysfunction (Ohkawa *et al.* 2013).

Taken together, the LGI1-ADAM22 complex functions as an antiepileptogenetic factor of which dysfunction causes an imbalance of excitation and inhibition, either too much of excitation or too little of inhibition, in all stages of brain development. This is the likely mechanism underlying intractable seizures in our patient. Notably, the majority of proteins involved in this pathway, ADAM22, LGI1, ADAM23, stargazin (CACNG2), AMPA and NMDA receptor subunits (**Figure 14**) are genetically linked to epilepsy in mice and/or humans (Letts *et al.* 1998; Kalachikov *et al.* 2002; Sagane *et al.* 2005; Owuor *et al.* 2009; Chabrol *et al.* 2010; Christie *et al.* 2010; Yu *et al.* 2010; Carvill *et al.* 2013b; Lemke *et al.* 2013; Lesca *et al.* 2013).

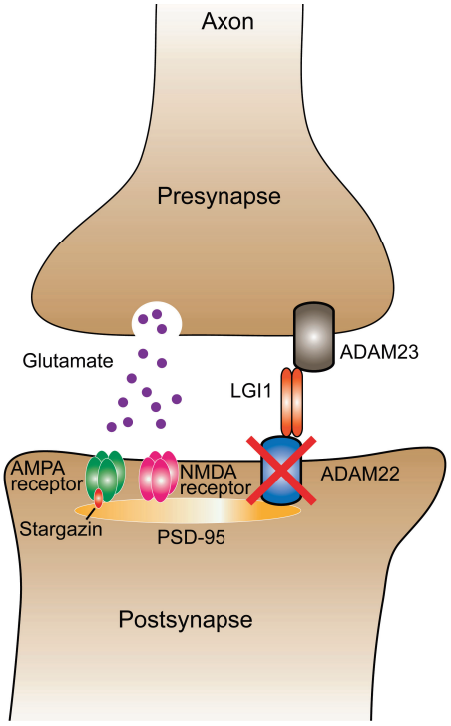


Figure 14 The synaptic LGI1-ADAM22/23 complex. LGI1 is a secreted glycoprotein, which binds to its transmembranic receptors ADAM23 and ADAM22 at the synapses. In this study, functional assays with ADAM22 mutants, which harbour changes seen in the patient, indicated that the ability of the mutants to bind to LGI1 and PSD-95 is abolished (red cross). The impairment of the LGI1-ADAM22/23 complex and its interaction with PSD-95 disrupts development and functional maturation of synapses, where AMPA- and NMDA-type glutamate receptors play an important role, and causes an imbalance of excitation and inhibition throughout life.

The patient with *ADAM22* variants is asymptomatic at birth but at 2-3 months of age, the disease progressed rapidly with recurrent, pharmacoresistant seizures. A remarkable feature of the disease is the rapidly progressing cerebral atrophy that became apparent two months after the disease onset. It is possible that the frequent seizure activity contributes to the cortical atrophy but ADAM22 dysfunction may also have a direct negative effect on neuronal development. The symptom onset of the patient is in line with the *Adam22* knockout mice, which develop lethal seizures at two postnatal weeks. These symptoms in our patient and in the mice comply not only with the proposed function of ADAM22 in the functional maturation of synapses (**Figure 14**), but also with the temporal expression pattern of *ADAM22*, which increases substantially at postnatal stages (III, **figure e-5**).

Our patient is alive at the age of 26 and the disease has remained stable after initial rapid progression. The cause for the less dramatic clinical course of our patient compared to *Adam22*$^{-/-}$ mice may be that the frameshift deletion occurs in exon 27, which is alternatively spliced and not expressed in

all known *ADAM22* isoforms (III, **figure e-4**). This exon is not present in the peripheral nervous system but expressed widely in the CNS, where its prenatal expression is lower compared to other *ADAM22* exons, but postnatally, it reaches the same level (Sagane *et al.* 2005; Özkaynak *et al.* 2010). In accordance with the absence of exon 27 from peripheral nervous system, *Adam22*$^{-/-}$ mice show defective myelination in peripheral neurons, which, however, appears to be normal in the patient. Only few other exons encoding cytoplasmic parts of ADAM22 are alternatively spliced and expressed only in the CNS. This may be one reason why *ADAM22*-associated encephalopathy appears to be very rare. It remains to be seen what the symptoms would be if both biallelic loss-of-function variants affect exons that are not alternatively spliced. Those variants could be embryonic lethal or lethal in early life as in the mouse model. In any case, identification of additional affected individuals with *ADAM22* variants would confirm that biallelic loss-of-function variants in *ADAM22* are associated with a disease involving severe encephalopathy with epilepsy.

6 CONCLUSIONS

We utilised WES to characterise genetic causes of severe epileptic syndromes in patients who had remained without molecular diagnosis prior to the study. We hypothesised that the study cohorts may be enriched for novel genetic findings with the potential to increase our understanding of biological mechanisms underlying syndromes with epilepsy.

Importantly, we identified two new, definite disease genes, *KCNC1* and *UBA5*, whose characterisation provides new insight on the molecular genetic basis of these syndromes affecting the CNS. In *KCNC1*, we unexpectedly characterised a single, recurrent *de novo* missense mutation, which explains a substantial proportion of unsolved sporadic or autosomal dominantly inherited PME cases. We termed the new worldwide PME subtype caused by the recurrent *KCNC1* mutation as MEAK. In *UBA5*, we described autosomal recessively inherited variants that cause a severe, infantile-onset encephalopathy. This *UBA5*-associated syndrome is likely to be present throughout populations of European descent, because the missense variant present in all five identified *UBA5* families is relatively frequent in Europeans, in Finns in particular. Finally, identification of *ADAM22* loss-of-function variants in a patient with rapidly progressing encephalopathy with seizures further implicates that LGI1-ADAM22/23 complex plays an important antiepileptogenic role in the synapses, a hypothesis originally derived from biochemical studies and mouse models. However, to confirm the causal role of ADAM22 dysfunction in human epilepsies, identification of additional patients is needed.

The three disease genes characterised in this study represent distinct molecular pathways. *KCNC1* encodes a potassium ion channel $K_V3.1$, which functions in the generation of high-frequency action potentials in the CNS. While ion channels underlie many epilepsy syndromes, $K_V3.1$ is the first one linked to PMEs. $K_V3.1$ is a key potassium channel in inhibitory interneurons and thus we hypothesise that reduced neuronal inhibition is the underlying epileptogenic mechanism. However, since MEAK patients also have ataxia, and many of them present with cognitive dysfunction and cerebellar atrophy, consequence of $K_V3.1$ dysfunction in the CNS is not limited to promoting seizure activity.

Disinhibition likely plays a role also in ADAM22-associated epilepsy, given that the LGI1-ADAM22 complex is expressed in inhibitory interneurons and its dysfunction has been shown to impair the excitatory input of the interneurons, leading to reduced inhibitory signalling. However, the impairment of LGI1-ADAM22 in excitatory neurons may lead to hyperexcitation, so likely a combination of hyperexcitation and disinhibition underlies the seizure phenotype when this complex is affected.

In study II, we determined that biallelic *UBA5* variants cause a partial but not complete attenuation of UFM1 activation by UBA5, which likely leads to

decreased number UFM1-target protein conjugates in a cell. Compared to $K_V3.1$ and ADAM22, the mechanism how UBA5 and UFM1-system dysfunction leads to impaired CNS development and more specifically, epileptogenic activity, is more challenging to decipher, mainly because the role of UFM1-conjugation in the CNS has not been characterised. One possibility is that the pathomechanism underlies the previously implicated role of UFM1-system in the protection from ER stress induced apoptosis.

For all the three genes implicated in this study, further functional work is required to address the disease mechanisms in the context of neuronal networks and whole brain. It is to be noted that in none of the three syndromes identified in this study symptoms are limited to epileptic seizures. This indicates that the dysfunctional proteins play a key role in molecular pathways that are fundamental for various aspects of the CNS development and function. Presence of a variety of CNS symptoms in patients with a single genetic defect complies with the emerging theme that in rare and common neurological, neurodevelopmental and neuropsychiatric diseases the underlying genetic loci and biological mechanisms are partially shared and do not always agree with the clinically determined disease entities.

Our findings highlight the power of WES not only in identification of new disease genes but also in being able to expand genotype-phenotype associations in previously established disease genes. In addition to the identification of *KCNC1* in study I, we broadened the variety of clinical features associated with pathogenic variants in known PME genes and in those originally linked to other related epilepsy or neurodegenerative disorders. The 'genotype-first' approach, enabled by WES, made it possible to obtain a specific diagnosis for 31% of individuals in the clinically highly heterogeneous PME cohort, which had been subjected to previous molecular testing in known disease genes. The majority of solved cases were sporadic, which represents another factor why traditional genetic approaches would not have been able to provide most if any of the diagnoses. Presence of novel and clinically unexpected genetic findings in our cohort supports the utilisation of WES or WGS over gene panels in clinically heterogeneous cohorts of epilepsy patients. Altogether, findings in the PME cohort improve molecular diagnostics, demonstrate the importance of *de novo* mutations and aid the search for potential therapeutic interventions in PMEs. The exome data of still unsolved cases also serves as resource to identify additional PME genes.

An important message of this study is the power of collaboration. In the PME study, an international clinical cooperation extending to over 25-year period enabled establishment of a sample collection large enough in size to make significant progress in understanding the genetic basis of this very rare disease entity. In study II, we collaborated with the nationwide UK-based DDD study to identify two additional patients with *UBA5* variants. The DDD study is an example of a project, which truly promotes disease gene identification by sharing its massive amount of genetic and phenotypic data

via DECIPHER database and the European Genome-phenome Archive. Individual research centres and even individual countries simply do not often have the numbers to match what is required to increase our knowledge of genetics of rare and common disorders. When isolated collections are combined, identification of new genes is greatly facilitated. Here, this was exemplified by the discovery of the globally occurring MEAK, which is locally rare because it is caused by a sporadic *de novo* mutation with no geographical hotspots. Collaboration is also important in other levels of genetics research. In functional work, we collaborated with groups having years of experience with the proteins of interest, which enabled efficient and high-quality testing of the pathogenicity of the newly identified variants.

7 CONCLUDING REMARKS AND FUTURE PROSPECTS

This thesis and work by others demonstrate that NGS technologies have revolutionised gene discovery in Mendelian disorders. It is easy to predict that WES and WGS have an increasingly important role in routine diagnostics in the upcoming years. Indeed, the importance of the genetics-driven approach in medicine is recognised, for example, by UK- and US-based funding bodies that have budgeted hundreds of millions of dollars altogether to perform WGS in both rare and common disorders.

The biggest obstacle to utilisation of NGS in gene diagnostics is not anymore in producing the sequence data, which is now fast and relatively inexpensive; however, bearing in mind that in resource-poor countries NGS methods are still largely inaccessible. Instead, the major challenges are in the informatics side. WGS in particular, which produces tens of gigabytes of data per genome, causes substantial expenses related to requirements for data storage and computing power. Improving interpretation of the clinical significance of genetic variation, both in coding and non-coding regions, is a major hurdle to overcome. Over the past years, it has become evident that the complex nature of the genomic variation not only makes it challenging to detect needles from the genomic haystack but it is equally easy to incorrectly consider truly benign variation as the cause of the disease. Misinterpretation of the significance of variants has major consequences to both genetic counselling and determination of appropriate treatment for patients. Therefore, best practices of variant interpretation, such as those suggested by ACMG (Richards *et al.* 2015), should be followed when analysing genetic data. One laudable effort to improve clinical interpretation of genetic variation is ClinGen, which is a resource aiming to facilitate sharing of genomic data as well as to curate gene-phenotype relationships and variant pathogenicity classifications (Rehm *et al.* 2015). Moreover, it aims to develop machine-learning methods to facilitate interpretation of the clinical impact of genetic variants.

The majority of exome-sequenced patients remain unsolved in most studies, including ours. WGS, which captures virtually all types of genetic variants, is a natural step forward and likely increases the number of diagnoses (Gilissen *et al.* 2014). Given the rarity of most genetic disorders, data sharing is also of high importance to increase diagnostic yield. As an example, Epilepsy Genetics Initiative aims to gather exomes of unsolved patients from collaborating partners and reanalyse them collectively. The results are then returned to referring clinicians.

It has been intriguing to observe how social media and improved genetic diagnostics have allowed patients and their families to become increasingly involved in the process of sharing knowledge about genetic disorders and enhancing precision medicine. Families with specific genetic disorders have

established organisations, which act as patient advocacy groups, increase awareness of the disorder, share information between the families, provide peer-support and raise funds for research. There are several patient organisations in epilepsy syndromes, and also in MEAK, the new subtype of PME identified in this study, families have been in contact with each other and shared their experiences with the disease. A foundation for MEAK research is also being established by one of the families. Importantly, families have also started to have an active role in trying to find the cause of their disease. WES or WGS often yields variants in a gene that has not been linked to human disease. There are now several examples where parents have not remained passive with this inconclusive situation. Instead, they use social media to try to reach other families or researchers, who have patients with variants in the same gene. In one example, parents used social media to find other patients and this led to establishment of a new genetic syndrome in short time (Enns *et al.* 2014). This highlights how important it is to connect isolated patients with ultra-rare conditions.

How can the increasing knowledge of genetics be translated into targeted therapies? In epilepsies and most other disease entities, it is not an easy task, given the wide spectrum of underlying molecular pathways. It is promising to see, however, that already now some genetic diagnoses can lead to targeted therapies in epilepsies, a topic which is covered in section 2.4.3. In general however, it has been estimated that only 5% of human genes are both druggable and related to human disease, which poses challenges for development of new therapies (Hopkins and Groom 2002; Cheng *et al.* 2007). Intriguingly, MEAK represents a genetic disorder where small-molecules targeting the defective protein have been discovered. As mentioned in section 5.1.3.2, activating compounds for the $K_V3.1$ potassium channel have been isolated, which is encouraging in terms of potential therapeutic interventions of MEAK where $K_V3.1$-mediated potassium currents are reduced. The effect of these compounds on the mutant $K_V3.1$ channel will be tested using *in vitro* systems and a knock-in mouse with the analogous recurrent mutation in *Kcnc1*.

In Mendelian disorders, which are due to defects in single genes, gene therapy is a promising therapeutic approach. In gene therapy, genetic material is inserted into patient cells with the aim to either knock out mutated copy of the gene, replace a mutated gene with a functional one or deliver a new gene into the cell. However, reaching adequate specificity and efficiency has been the major obstacle of gene therapy. A highly promising tool to increase specificity is CRISPR-Cas9 genome editing tool, which is likely to become one of the most influential technologies in the history of molecular biology. It has been shown to be a precise and simple tool to make specific changes in DNA. It can be used, for example, to inactivate a gene by making a deletion in a specific site or to correct DNA sequence. Recently, CRISPR-Cas9 system has been successfully delivered to mouse neurons *in vivo* using adeno-associated viral vectors with high efficiency (Swiech *et al.*

2015). More compact and more precise versions of the CRISPR-Cas9 tool have since been introduced, providing increasing hope for therapeutic utility of this approach in a variety of genetic disorders (Ran *et al.* 2015a; Slaymaker *et al.* 2016). Adeno-associated viral vectors can be used to deliver genetic material into human CNS but the efficiency of this method to transduce target cells needs still improvement (Bourdenx *et al.* 2014). Yet, CRISPR-Cas9 system represents a highly tempting therapeutic approach in MEAK and other genetic disorders to correct the genetic defect. The utility of this approach could be tested in the knock-in mouse model of MEAK.

8 ACKNOWLEDGEMENTS

This study was carried out at the Institute for Molecular Medicine Finland FIMM, Molecular Neurology Research Program and Neuroscience Center of University of Helsinki as well as at the Folkhälsan Institute of Genetics, Helsinki, Finland. I wish to acknowledge Directors of FIMM Olli Kallioniemi and Jaakko Kaprio, Head of the Molecular Neurology Research Program Anu Wartiovaara, Director of Neuroscience Center Eero Castrén and Director of Folkhälsan Institute of Genetics Anna-Elina Lehesjoki for providing excellent infrastructure for research.

This thesis was funded by Doctoral Programme in Biomedicine, Emil Aaltonen Foundation, University of Helsinki Funds, Epilepsy Research Foundation, Arvo and Lea Ylppö Foundation, Finnish Brain Foundation, Paulo Foundation and Biomedicum Helsinki Foundation, all of which are sincerely acknowledged.

I have had the privilege of having two inspiring and supportive thesis supervisors. Professor Aarno Palotie, who has served as my supervisor from the beginning, has always pushed me to the right direction, selected interesting and high-level thesis projects and connected me with the best possible collaborators. The energetic and encouraging personality Aarno has is unrivalled and has had an important impact in particular in the early stages of my projects when the results were not yet that promising. I want to thank Professor Anna-Elina Lehesjoki for taking me as a member of also her group, which was an important step in my PhD studies as it enabled me to work with the excellent exome sequencing projects utilising unique patient materials. We developed a dynamic and productive collaboration and it was a great pleasure to work together with the projects and manuscripts.

I wish to greatly acknowledge late Academician of Science, Leena Peltonen-Palotie whose group I joined in 2009. I am grateful for her to have been able to start my career in human genetics in the best possible environment. I would like to thank the pre-examiners of my thesis, Helen Firth and Marjo Kestilä, for accepting the task in short notice and taking the time to evaluate my thesis. I want to thank for the constructive feedback and expertise in reviewing the thesis. I wish to acknowledge Professor Joris Veltman for accepting the role of serving as the opponent of my defence. It is a great honour to have such an expert in the field of human genetics to review my work. I also want to thank Professor Päivi Peltomäki for accepting the role of custos in the defence, and the members of my thesis committee, Professor Hannes Lohi and Päivi Saavalainen, for examining my thesis work over the years and providing constructive feedback.

I am fortunate to have been able to work in inspiring and high quality research projects with a number of collaborators. First, I wish to thank Professor Sam Berkovic for his efforts in establishing the worldwide network of PME researchers and clinicians. It has been a great privilege and joy to

analyse the genetic data from this unique cohort. I want to thank Karen Oliver for the excellent work in processing the clinical data of PME patients and pursuing the clinical assessment of MEAK further. I also want to thank Associate Professor Leanne Dibbens and Marta Bayly for the collaboration in the genetics part of the PME project as well as Professor Mark Daly and Kaitlin Samocha for providing statistical insight to the PME project.

We all know how demanding it is to do functional work in the lab. Therefore, I am deeply grateful for the numerous collaborators who have performed lab experiments in these projects. I would like to acknowledge Professor Holger Lerche, Snezana Maljevic and Professor Steven Petrou for performing electrophysiological work in the PME project. Professor Masaaki Komatsu and his lab members, Ryosuke Ishimura and Yoshinobu Ichimura in particular, are thanked for conducting numerous biochemical experiments in the *UBA5* project. Also, I would like to thank Satoshi Waguri for his group's contribution to the project. Masaki and Yuko Fukata are acknowledged for performing functional work in the *ADAM22* project so efficiently and with high standards. Finally, I would like to thank all other co-authors, clinicians, families and patients who have participated in the studies.

It has always been fun to come to the office. This is thanks to the positive atmosphere and the awesome people we have had in the FIMM Human Genomics department. I want to thank my colleagues for the enjoyable time and (non)scientific discussions in and outside the office. Thank you Tiia, Olli, Luca, Meri, Taru, Annu, Jaakko, Antti, Ida, Emmi, Anni, Perttu, Priit, Johannes, Peter W×2, Matti, Marjo, Iiro, Tero, Annika, Marjis, Virpi, Ansku, Hanski, Alfredo, Vishal, Liisa U., Diana, Liisa T, Marine, Sini, Pyry, Paavo, Himanshu, Christian, Anna, Kimmo, Pietari, Joel, Elmo, P-P, Jarkko, Huei-yi. You and other members of the human genetics group over the years will be missed! A special thanks to members of the autism group, Karola (thank you for supervising my Master's thesis and writing several letters of recommendation), Helena and Mari, who guided me after joining FIMM in 2009. I also want to thank senior members of the Human Genomics department: Professor Samuli Ripatti, William Hennah, Maija Wessman, Elisabeth Widen, Janna Saarela, Hannele Laivuori. Thanks also to Aarno's group members in Sanger: Padhraig, Bendik (thanks for accommodating me!), Verneri, Priit, Chloe, Lucy, Alison, Christine.

I want thank Sari Kivikko for taking care of all the administrative issues over the years. Sari is a person who can be trusted 100% and has the incredible ability to organise everything so efficiently. Similarly, in the lab, Eija Hämäläinen and Paula Hakala have kept the samples and projects organised and I want to thank you for your precise and trustworthy work!

I wish to thank all members of the AEL group. In particular, I want to thank Anni Laari who has worked hard in the lab in the PEHO- and PEHO-like projects. I cannot thank you enough for performing the lab work that I have passed on your shoulders. I also want to thank Tarja Joensuu, who has

been invaluable in the PME projects, and Anna-Kaisa Anttonen and Tarja Linnankivi who have been critical for the clinical parts of the projects. Thank you Salla for your work in the PME project, and all others in the group: Sinikka, Marjatta, Inken, Carolina, Kata, Edu, Otto, Teija, Ann-Liz, Olesya and Saara.

It has been important to have good friends to spend time with outside the office. First, I want to thank the Bioboys – Heikki, Manu, JP and Valtteri – with whom I have had much fun in Finland and abroad since the beginning of our studies in the university. In particular, I want to thank Heikki for providing peer support in PhD studies at FIMM. Bioboys, maybe one day we will have the company that we have dreamed about establishing. We only need a good business plan! Maybe that will be decided in the road trip in the autumn? Another group of people I want to thank for the enjoyable time spent together over the years is the Proffakerho (Professor Club), which is named after our nerdy behaviour in the school. Jouko, Onni, Häkä, Saku, Jyrki, VP, Niko, thanks for your company in the past >15 years and let's have more of the fun nerdy stuff in the future too! Now one of us is closer of becoming an actual Professor!

Finally, I want to thank my family for providing the best possible environment to grow up. Thank you mum and dad, Mirja and Pekka, for the support and encouragement! I wish to thank my sister Riikka and her husband Sami and their two lovely sons Matias and Niklas for the good times together. And, to my dear wife Katriina: thank you for the love and support during these years. I am looking forward to our future adventures together!

Helsinki, 23.5.2016

Mikko Muona

9 REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010). A method and server for predicting damaging missense mutations. *Nat Methods* 7:248-249.

Afawi Z, Mandelstam S, Korczyn AD, Kivity S, Walid S, Shalata A, Oliver KL, Corbett M, Gecz J, Berkovic SF, Jackson GD (2013). TBC1D24 mutation associated with focal epilepsy, cognitive impairment and a distinctive cerebro-cerebellar malformation. *Epilepsy Res* 105:240-244.

Aggarwal SK, MacKinnon R (1996). Contribution of the S4 Segment to Gating Charge in the Shaker K+ Channel. *Neuron* 16:1169-1177.

Akawi N, McRae J, Ansari M, Balasubramanian M, Blyth M, Brady AF, Clayton S, Cole T, Deshpande C, Fitzgerald TW, Foulds N, Francis R, Gabriel G, Gerety SS, Goodship J, Hobson E, Jones WD, Joss S, King D, Klena N, Kumar A, Lees M, Lelliott C, Lord J, McMullan D, O'Regan M, Osio D, Piombo V, Prigmore E, Rajan D, Rosser E, Sifrim A, Smith A, Swaminathan GJ, Turnpenny P, Whitworth J, Wright CF, Firth HV, Barrett JC, Lo CW, FitzPatrick DR, Hurles ME, the DDD study (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat Genet* 47:1363-1369.

Albert FW, Kruglyak L (2015). The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16:197-212.

Alvaro G, Marasco A (2013). Hydantoin derivatives useful as kv3 inhibitors, Google Patents.

Annegers JF, Hauser WA, Anderson VE, Kurland LT (1982). The risks of seizure disorders among relatives of patients with childhood onset epilepsy. *Neurology* 32:174-9.

Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, Wang J, Wu M, Liu X, Tian G, Wang J, Wang J, Yang H, Zhang X (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* 12:R95.

Atkin G, Paulson H (2014). Ubiquitin Pathways in Neurodegenerative Disease. *Front Mol Neurosci* 7.

Bacik J-P, Walker JR, Ali M, Schimmer AD, Dhe-Paganon S (2010). Crystal Structure of the Human Ubiquitin-activating Enzyme 5 (UBA5) Bound to ATP: Mechanistic Insights Into a Minimalistic E1 Enzyme. *J Biol Chem* 285:20273-20280.

Baets J, Deconinck T, Smets K, Goossens D, Van den Bergh P, Dahan K, Schmedding E, Santens P, Rasic VM, Van Damme P, Robberecht W, De Meirleir L, Michielsens B, Del-Favero J, Jordanova A, De Jonghe P (2010). Mutations in SACS cause atypical and late-onset forms of ARSACS. *Neurology* 75:1181-1188.

Baraban SC, Dinday MT, Hortopan GA (2013). Drug screening in Scn1a zebrafish mutant identifies clemizole as a potential Dravet syndrome treatment. *Nat Commun* 4:2410.

Barcia G, Fleming MR, Deligniere A, Gazula VR, Brown MR, Langouet M, Chen H, Kronengold J, Abhyankar A, Cilio R, Nitschke P, Kaminska A, Boddaert N, Casanova JL, Desguerre I, Munnich A, Dulac O, Kaczmarek LK, Colleaux L, Nabbout R (2012). De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. *Nat Genet* 44:1255-9.

Baron M (2001). The search for complex disease genes: fault by linkage or fault by association? *Mol Psychiatry* 6:143-9.

Baulac S, Ishida S, Marsan E, Miquel C, Biraben A, Nguyen DK, Nordli D, Cossette P, Nguyen S, Lambrecq V, Vlaicu M, Daniau M, Bielle F, Andermann E, Andermann F, Leguern E, Chassoux F, Picard F (2015). Familial focal epilepsy with focal cortical dysplasia due to DEPDC5 mutations. *Ann Neurol* 77:675-683.

Bearden D, Strong A, Ehnot J, DiGiovine M, Dlugos D, Goldberg EM (2014). Targeted treatment of migrating partial seizures of infancy with quinidine. *Ann Neurol* 76:457-461.

Bellus GA, Gaudenz K, Zackai EH, Clarke LA, Szabo J, Francomano CA, Muenke M (1996). Identical mutations in three different fibroblast growth factor receptor genes in autosomal dominant craniosynostosis syndromes. *Nat Genet* 14:174-176.

Benson G (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573-580.

Berg AT, Berkovic SF, Brodie MJ, Buchhalter J, Cross JH, Van Emde Boas W, Engel J, French J, Glauser TA, Mathern GW, Moshé SL, Nordli D, Plouin P, Scheffer IE (2010). Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005–2009. *Epilepsia* 51:676-685.

Berkovic SF, Andermann F, Carpenter S, Wolfe LS (1986). Progressive Myoclonus Epilepsies: Specific Causes and Diagnosis. *N Engl J Med* 315:296-305.

Berkovic SF, Howell RA, Hay DA, Hopper JL (1998). Epilepsies in twins: genetics of the major epilepsy syndromes. *Ann Neurol* 43:435-45.

Bianca M, Bianca S, Vecchio I, Raffaele R, Ingegnosi C, Nicoletti F (2003). Gerstmann–Sträussler–Scheinker disease with P102L–V129 mutation: a case with psychiatric manifestations at onset. *Ann Genet* 46:467-469.

Boillot M, Huneau C, Marsan E, Lehongre K, Navarro V, Ishida S, Dufresnois B, Ozkaynak E, Garrigue J, Miles R, Martin B, Leguern E, Anderson MP, Baulac S (2014). Glutamatergic neuron-targeted loss of LGI1 epilepsy gene results in seizures. *Brain* 137:2984-2996.

Bourdenx M, Dutheil N, Bezard E, Dehay B (2014). Systemic gene delivery to the central nervous system using Adeno-associated virus. *Front Mol Neurosci* 7.

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* 14:681-691.

Brown P, Cathala F, Castaigne P, Gajdusek DC (1986). Creutzfeldt-Jakob disease: clinical analysis of a consecutive series of 230 neuropathologically verified cases. *Ann Neurol* 20:597-602.

Buzzi A, Chikhladze M, Falcicchia C, Paradiso B, Lanza G, Soukupova M, Marti M, Morari M, Franceschetti S, Simonato M (2012). Loss of cortical GABA terminals in Unverricht-Lundborg disease. *Neurobiol Dis* 47:216-224.

Cai Y, Pi W, Sivaprakasam S, Zhu X, Zhang M, Chen J, Makala L, Lu C, Wu J, Teng Y, Pace B, Tuan D, Singh N, Li H (2015). UFBP1, a key component of the Ufm1 conjugation system, is essential for ufmylation-mediated regulation of erythroid development. *PLoS Genet* 11:e1005643.

Campbell CD, Eichler EE (2013). Properties and rates of germline mutations in humans. *Trends Genet* 29:575-584.

Campeau PM, Kasperaviciute D, Lu JT, Burrage LC, Kim C, Hori M, Powell BR, Stewart F, Felix TM, van den Ende J, Wisniewska M, Kayserili H, Rump P, Nampoothiri S, Aftimos S, Mey A, Nair LD, Begleiter ML, De Bie I, Meenakshi G, Murray ML, Repetto GM, Golabi M, Blair E, Male A, Giuliano F, Kariminejad A, Newman WG, Bhaskar SS, Dickerson JE, Kerr B, Banka S, Giltay JC, Wieczorek D, Tostevin A, Wiszniewska J, Cheung SW, Hennekam RC, Gibbs RA, Lee BH, Sisodiya SM (2014). The genetic basis of DOORS syndrome: an exome-sequencing study. *Lancet Neurol* 13:44-58.

Canafoglia L, Robbiano A, Pareyson D, Panzica F, Nanetti L, Giovagnoli AR, Venerando A, Gellera C, Franceschetti S, Zara F (2014). Expanding sialidosis spectrum by genome-wide screening: NEU1 mutations in adult-onset myoclonus. *Neurology* 82:2003-2006.

Cannon SC (2010). Voltage-sensor mutations in channelopathies of skeletal muscle. *J Physiol* 588:1887-1895.

Cardarelli WJ, Smith BJ (2010). The burden of epilepsy to patients and payers. *Am J Manag Care* 16:S331-6.

Carvill GL, Heavin SB, Yendle SC, McMahon JM, O'Roak BJ, Cook J, Khan A, Dorschner MO, Weaver M, Calvert S, Malone S, Wallace G, Stanley T, Bye AME, Bleasel A, Howell KB, Kivity S, Mackay MT, Rodriguez-Casero V, Webster R, Korczyn A, Afawi Z,

Zelnick N, Lerman-Sagie T, Lev D, Moller RS, Gill D, Andrade DM, Freeman JL, Sadleir LG, Shendure J, Berkovic SF, Scheffer IE, Mefford HC (2013a). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat Genet* 45:825-830.

Carvill GL, Regan BM, Yendle SC, O'Roak BJ, Lozovaya N, Bruneau N, Burnashev N, Khan A, Cook J, Geraghty E, Sadleir LG, Turner SJ, Tsai M-H, Webster R, Ouvrier R, Damiano JA, Berkovic SF, Shendure J, Hildebrand MS, Szepetowski P, Scheffer IE, Mefford HC (2013b). GRIN2A mutations cause epilepsy-aphasia spectrum disorders. *Nat Genet* 45:1073-1076.

Chabrol E, Navarro V, Provenzano G, Cohen I, Dinocourt C, Rivaud-Pechoux S, Fricker D, Baulac M, Miles R, Leguern E, Baulac S (2010). Electroclinical characterization of epileptic seizures in leucine-rich, glioma-inactivated 1-deficient mice. *Brain* 133:2749-62.

Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 25:71-5.

Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, Coban Akdemir ZH, Doheny K, Scott AF, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling H, Hetrick K, Watkins L, Patterson KE, Reinier F, Blue E, Muzny D, Kircher M, Bilguvar K, López-Giráldez F, Sutton VR, Tabor HK, Leal SM, Gunel M, Mane S, Gibbs RA, Boerwinkle E, Hamosh A, Shendure J, Lupski JR, Lifton RP, Valle D, Nickerson DA, Bamshad MJ (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* 97:199-215.

Christie LA, Russell TA, Xu J, Wood L, Shepherd GM, Contractor A (2010). AMPA receptor desensitization mutation results in severe developmental phenotypes and early postnatal lethality. *Proc Natl Acad Sci U S A* 107:9412-7.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6:80-92.

Claes L, Del-Favero J, Ceulemans B, Lagae L, Van Broeckhoven C, De Jonghe P (2001). De Novo Mutations in the Sodium-Channel Gene SCN1A Cause Severe Myoclonic Epilepsy of Infancy. *Am J Hum Genet* 68:1327-1332.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* 104:19428-19433.

Cooper DN, Chen J-M, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD (2010a). Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mutat* 31:631-655.

Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA (2010b). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Meth* 7:250-251.

Corbett MA, Bahlo M, Jolly L, Afawi Z, Gardner AE, Oliver KL, Tan S, Coffey A, Mulley JC, Dibbens LM, Simri W, Shalata A, Kivity S, Jackson GD, Berkovic SF, Gecz J (2010). A Focal Epilepsy and Intellectual Disability Syndrome Is Due to a Mutation in TBC1D24. *Am J Hum Genet* 87:371-375.

Cutting GR (2015). Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet* 16:45-56.

D'Alessandro LCA, Al Turki S, Manickaraj AK, Manase D, Mulder BJM, Bergin L, Rosenberg HC, Mondal T, Gordon E, Lougheed J, Smythe J, Devriendt K, Bhattacharya S, Watkins H, Bentham J, Bowdin S, Hurles ME, Mital S (2015). Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect. *Genet Med*.

de Boer HM, Mula M, Sander JW (2008). The global burden and stigma of epilepsy. *Epilepsy Behav* 12:540-546.

de Goede C, Yue WW, Yan G, Ariyaratnam S, Chandler KE, Downes L, Khan N, Mohan M, Lowe M, Banka S (2016). Role of reverse phenotyping in interpretation of next generation sequencing data and a review of INPP5E related disorders. *Eur J Paediatr Neurol* 20:286-295.

de Kovel CGF, Trucks H, Helbig I, Mefford HC, Baker C, Leu C, Kluck C, Muhle H, von Spiczak S, Ostertag P, Obermeier T, Kleefuß-Lie AA, Hallmann K, Steffens M, Gaus V, Klein KM, Hamer HM, Rosenow F, Brilstra EH, Kasteleijn-Nolst Trenité D, Swinkels MEM, Weber YG, Unterberger I, Zimprich F, Urak L, Feucht M, Fuchs K, Møller RS, Hjalgrim H, De Jonghe P, Suls A, Rückert I-M, Wichmann H-E, Franke A, Schreiber S, Nürnberg P, Elger CE, Lerche H, Stephani U, Koeleman BPC, Lindhout D, Eichler EE, Sander T (2010). Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* 133:23-32.

Delemotte L, Treptow W, Klein ML, Tarek M (2010). Effect of Sensor Domain Mutations on the Properties of Voltage-Gated Ion Channels: Molecular Dynamics Studies of the Potassium Channel Kv1.2. *Biophys J* 99:L72-L74.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491-498.

Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*.

Dibbens LM, de Vries B, Donatello S, Heron SE, Hodgson BL, Chintawar S, Crompton DE, Hughes JN, Bellows ST, Klein KM, Callenbach PMC, Corbett MA, Gardner AE, Kivity S, Iona X, Regan BM, Weller CM, Crimmins D, O'Brien TJ, Guerrero-Lopez R, Mulley JC, Dubeau F, Licchetta L, Bisulli F, Cossette P, Thomas PQ, Gecz J, Serratosa J, Brouwer OF, Andermann F, Andermann E, van den Maagdenberg AMJM, Pandolfo M, Berkovic SF, Scheffer IE (2013). Mutations in DEPDC5 cause familial focal epilepsy with variable foci. *Nat Genet* 45:546-551.

Do R, Kathiresan S, Abecasis GR (2012). Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* 21:R1-R9.

Duan R, Shi Y, Yu L, Zhang G, Li J, Lin Y, Guo J, Wang J, Shen L, Jiang H, Wang G, Tang B (2016). UBA5 mutations cause a new form of autosomal recessive cerebellar ataxia. *PLoS ONE* 11:e0149039.

Ducros A, Denier C, Joutel A, Vahedi K, Michel A, Darcel F, Madigand M, Guerouaou D, Tison F, Julien J, Hirsch E, Chedru F, Bisgård C, Lucotte G, Després P, Billard C, Barthez MA, Ponsot G, Bousser MG, Tournier-Lasserve E (1999). Recurrence of the T666M Calcium Channel CACNA1A Gene Mutation in Familial Hemiplegic Migraine with Progressive Cerebellar Ataxia. *Am J Hum Genet* 64:89-98.

Enns GM, Shashi V, Bainbridge M, Gambello MJ, Zahir FR, Bast T, Crimian R, Schoch K, Platt J, Cox R, Bernstein JA, Scavina M, Walter RS, Bibb A, Jones M, Hegde M, Graham BH, Need AC, Oviedo A, Schaaf CP, Boyle S, Butte AJ, Chen R, Clark MJ, Haraksingh R, Cowan TM, He P, Langlois S, Zoghbi HY, Snyder M, Gibbs RA, Freeze HH, Goldstein DB (2014). Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genet Med* 16:751-758.

Epi4K Consortium, Epilepsy Phenome/Genome Project (2013). De novo mutations in epileptic encephalopathies. *Nature* 501:217-221.

EpiPM Consortium (2015). A roadmap for precision medicine in the epilepsies. *Lancet Neurol* 14:1219-1228.

Epstein DJ (2009). Cis-regulatory mutations in human disease. *Brief Funct Genomic Proteomic* 8:310-316.

Eriksson M, Brown WT, Gordon LB, Glynn MW, Singer J, Scott L, Erdos MR, Robbins CM, Moses TY, Berglund P, Dutra A, Pak E, Durkin S, Csoka AB, Boehnke M, Glover TW, Collins FS (2003). Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* 423:293-298.

Escayg A, MacDonald BT, Meisler MH, Baulac S, Huberfeld G, An-Gourfinkel I, Brice A, LeGuern E, Moulard B, Chaigne D, Buresi C, Malafosse A (2000). Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+2. *Nat Genet* 24:343-345.

Espinosa F, McMahon A, Chan E, Wang S, Ho CS, Heintz N, Joho RH (2001). Alcohol Hypersensitivity, Increased Locomotion, and Spontaneous Myoclonus in Mice Lacking the Potassium Channels Kv3.1 and Kv3.3. *J Neurosci* 21:6657-6665.

EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project, Epi4K Consortium (2014). De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am J Hum Genet* 95:360-370.

European Comission. (2016). "Rare Diseases." Retrieved Jan 21, from http://ec.europa.eu/health/rare_diseases/policy/index_en.htm.

Everett CM, Wood NW (2004). Trinucleotide repeats and neurodegenerative disease. *Brain* 127:2385-2405.

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet*.

Falace A, Filipello F, La Padula V, Vanni N, Madia F, De Pietri Tonelli D, de Falco FA, Striano P, Dagna Bricarelli F, Minetti C, Benfenati F, Fassio A, Zara F (2010). TBC1D24, an ARF6-Interacting Protein, Is Mutated in Familial Infantile Myoclonic Epilepsy. *Am J Hum Genet* 87:365-370.

Fang H, Wu Y, Narzisi G, O'Rawe JA, Barron LT, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC, Lyon GJ (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med* 6:89.

Feenstra B, Pasternak B, Geller F, Carstensen L, Wang T, Huang F, Eitson JL, Hollegaard MV, Svanstrom H, Vestergaard M, Hougaard DM, Schoggins JW, Jan LY, Melbye M, Hviid A (2014). Common variants associated with general and MMR vaccine-related febrile seizures. *Nat Genet* 46:1274-1282.

Figueroa KP, Minassian NA, Stevanin G, Waters M, Garibyan V, Forlani S, Strzelczyk A, Bürk K, Brice A, Dürr A, Papazian DM, Pulst SM (2010). KCNC3: phenotype, mutations, channel biophysics—a study of 260 familial ataxia patients. *Hum Mutat* 31:191-196.

Figueroa KP, Waters MF, Garibyan V, Bird TD, Gomez CM, Ranum LPW, Minassian NA, Papazian DM, Pulst SM (2011). Frequency of KCNC3 DNA Variants as Causes of Spinocerebellar Ataxia 13 (SCA13). *PLoS ONE* 6:e17811.

Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Vooren SV, Moreau Y, Pettett RM, Carter NP (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* 84:524-533.

Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, Engel J, Forsgren L, French JA, Glynn M, Hesdorffer DC, Lee BI, Mathern GW, Moshé SL, Perucca E, Scheffer IE, Tomson T, Watanabe M, Wiebe S (2014). ILAE Official Report: A practical clinical definition of epilepsy. *Epilepsia* 55:475-482.

Fisher RS, van Emde Boas W, Blume W, Elger C, Genton P, Lee P, Engel J, Jr. (2005). Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia* 46:470-2.

Flicek P, Birney E (2009). Sense from sequence reads: methods for alignment and assembly. *Nat Meth* 6:S6-S12.

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT (2011). LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* 32:557-563.

Franceschetti S, Michelucci R, Canafoglia L, Striano P, Gambardella A, Magaudda A, Tinuper P, La Neve A, Ferlazzo E, Gobbi G, Giallonardo AT, Capovilla G, Visani E, Panzica F, Avanzini G, Tassinari CA, Bianchi A, Zara F (2014). Progressive myoclonic epilepsies: definitive and still undetermined causes. *Neurology* 82:405-411.

Fromer M, Moran Jennifer L, Chambert K, Banks E, Bergen Sarah E, Ruderfer Douglas M, Handsaker Robert E, McCarroll Steven A, O Donovan Michael C, Owen Michael J,

Kirov G, Sullivan Patrick F, Hultman Christina M, Sklar P, Purcell Shaun M (2012). Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. *Am J Hum Genet* 91:597-607.

Fuentes Fajardo KV, Adams D, Program NCS, Mason CE, Sincan M, Tifft C, Toro C, Boerkoel CF, Gahl W, Markello T (2012). Detecting false-positive signals in exome sequencing. *Hum Mutat* 33:609-613.

Fukata Y, Adesnik H, Iwanaga T, Bredt DS, Nicoll RA, Fukata M (2006). Epilepsy-related ligand/receptor complex LGI1 and ADAM22 regulate synaptic transmission. *Science* 313:1792-1795.

Fukata Y, Lovero KL, Iwanaga T, Watanabe A, Yokoi N, Tabuchi K, Shigemoto R, Nicoll RA, Fukata M (2010). Disruption of LGI1–linked synaptic complex causes abnormal synaptic transmission and epilepsy. *Proc Natl Acad Sci U S A* 107:3799-3804.

Gan L, Kaczmarek LK (1998). When, where, and how much? Expression of the Kv3.1 potassium channel in high-frequency firing neurons. *J Neurobiol* 37:69-79.

GATK Best Practices. (2016a). "Should I analyze my samples alone or together?" Retrieved Jan 26, 2016, from http://gatkforums.broadinstitute.org/gatk/discussion/4150/should-i-analyze-my-samples-alone-or-together.

GATK Best Practices. (2016b). "Variant Quality Score Recalibration (VQSR)." Retrieved Jan 26, 2016, from http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr.

Gazal S, Sahbatou M, Babron MC, Genin E, Leutenegger AL (2014). FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics* 30:1940-1.

George AL (2004). Inherited Channelopathies Associated with Epilepsy. *Epilepsy Curr* 4:65-70.

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R, Tearle R, Bo T, Pfundt R, Yntema HG, de Vries BBA, Kleefstra T, Brunner HG, Vissers LELM, Veltman JA (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511:344-347.

Gilissen C, Hoischen A, Brunner H, Veltman J (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biol* 12:228.

Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012). Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 20:490-497.

Glaser RL, Morison IM (2009). Equality of the Sexes? Parent-of-Origin Effects on Transcription and de novo Mutations. Bioinformatics for Systems Biology. S. Krawetz, Humana Press.

Global Genes. (2016). "Global genes website "RARE Facts and Statistics." https://globalgenes.org/rare-diseases-facts-statistics/ " Accessed January 20, 2016."

Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, Sunyaev S (2013). Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* 14:460-470.

Goriely A, McVean GA, Rojmyr M, Ingemarsson B, Wilkie AO (2003). Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science* 301:643-6.

Goto Y, Nonaka I, Horai S (1990). A mutation in the tRNA(Leu)(UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nature* 348:651-3.

Gu Y, Barry J, McDougel R, Terman D, Gu C (2012). Alternative Splicing Regulates Kv3.1 Polarized Targeting to Adjust Maximal Spiking Frequency. *J Biol Chem* 287:1755-1769.

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, Young AB, Shoulson I, Bonilla E, Martin JB (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306:234-238.

Guven A, Tolun A (2013). TBC1D24 truncating mutation resulting in severe neurodegeneration. *J Med Genet* 50:199-202.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33:D514-D517.

Harkin LA, McMahon JM, Iona X, Dibbens L, Pelekanos JT, Zuberi SM, Sadleir LG, Andermann E, Gill D, Farrell K, Connolly M, Stanley T, Harbord M, Andermann F, Wang J, Batish SD, Jones JG, Seltzer WK, Gardner A, The Infantile Epileptic Encephalopathy Referral Consortium, Sutherland G, Berkovic SF, Mulley JC, Scheffer IE (2007). The spectrum of SCN1A-related infantile epileptic encephalopathies. *Brain* 130:843-852.

Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, Franke A, Muhle H, de Kovel C, Baker C, von Spiczak S, Kron KL, Steinich I, Kleefusz-Lie AA, Leu C, Gaus V, Schmitz B, Klein KM, Reif PS, Rosenow F, Weber Y, Lerche H, Zimprich F, Urak L, Fuchs K, Feucht M, Genton P, Thomas P, Visscher F, de Haan G-J, Moller RS, Hjalgrim H, Luciano D, Wittig M, Nothnagel M, Elger CE, Nurnberg P, Romano C, Malafosse A, Koeleman BPC, Lindhout D, Stephani U, Schreiber S, Eichler EE, Sander T (2009). 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat Genet* 41:160-162.

Helbig KL, Farwell Hagman KD, Shinde DN, Mroske C, Powis Z, Li S, Tang S, Helbig I (2016). Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. *Genet Med*.

Heron SE, Scheffer IE, Iona X, Zuberi SM, Birch R, McMahon JM, Bruce CM, Berkovic SF, Mulley JC (2010). De novo SCN1A mutations in Dravet syndrome and related epileptic encephalopathies are largely of paternal origin. *J Med Genet* 47:137-141.

Heron SE, Smith KR, Bahlo M, Nobili L, Kahana E, Licchetta L, Oliver KL, Mazarib A, Afawi Z, Korczyn A, Plazzi G, Petrou S, Berkovic SF, Scheffer IE, Dibbens LM (2012). Missense mutations in the sodium-gated potassium channel gene KCNT1 cause severe autosomal dominant nocturnal frontal lobe epilepsy. *Nat Genet* 44:1188-90.

Hesdorffer DC, Logroscino G, Benn EKT, Katri N, Cascino G, Hauser WA (2011). Estimating risk for developing epilepsy: A population-based study in Rochester, Minnesota. *Neurology* 76:23-27.

Hicks S, Wheeler DA, Plon SE, Kimmel M (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum Mutat* 32:661-668.

Hildebrand MS, Dahl H-HM, Damiano JA, Smith RJH, Scheffer IE, Berkovic SF (2013). Recent advances in the molecular genetics of epilepsy. *J Med Genet* 50:271-279.

Ho CS, Grange RW, Joho RH (1997). Pleiotropic effects of a disrupted K+ channel gene: Reduced body weight, impaired motor skill and muscle contraction, but no seizures. *Proc Natl Acad Sci U S A* 94:1533-1538.

Holt IJ, Harding AE, Petty RK, Morgan-Hughes JA (1990). A new mitochondrial disease associated with mitochondrial DNA heteroplasmy. *Am J Hum Genet* 46:428-33.

Hopkins AL, Groom CR (2002). The druggable genome. *Nat Rev Drug Discov* 1:727-730.

Hsiao K, Baker HF, Crow TJ, Poulter M, Owen F, Terwilliger JD, Westaway D, Ott J, Prusiner SB (1989). Linkage of a prion protein missense variant to Gerstmann-Straussler syndrome. *Nature* 338:342-345.

Huber C, Dias-Santagata D, Glaser A, O'Sullivan J, Brauner R, Wu K, Xu X, Pearce K, Wang R, Uzielli MLG, Dagoneau N, Chemaitilly W, Superti-Furga A, Santos HD, Megarbane A, Morin G, Gillessen-Kaesbach G, Hennekam R, Burgt IVd, Black GCM, Clayton PE, Read A, Merrer ML, Scambler PJ, Munnich A, Pan Z-Q, Winter R, Cormier-Daire V (2005). Identification of mutations in CUL7 in 3-M syndrome. *Nat Genet* 37:1119-1124.

Hübner CA, Jentsch TJ (2002). Ion channel diseases. *Hum Mol Genet* 11:2435-2445.

Iglesias A, Anyane-Yeboa K, Wynn J, Wilson A, Truitt Cho M, Guzman E, Sisson R, Egan C, Chung WK (2014). The usefulness of whole-exome sequencing in routine clinical practice. *Genet Med* 16:922-931.

ILAE Consortium on Complex Epilepsies (2014). Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol* 13:893-903.
International HapMap Consortium (2003). The International HapMap Project. *Nature* 426:789-96.
International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945.
Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* advance online publication.
Irani SR, Alexander S, Waters P, Kleopa KA, Pettingill P, Zuliani L, Peles E, Buckley C, Lang B, Vincent A (2010). Antibodies to Kv1 potassium channel-complex proteins leucine-rich, glioma inactivated 1 protein and contactin-associated protein-2 in limbic encephalitis, Morvan's syndrome and acquired neuromyotonia. *Brain* 133:2734-2748.
Ishida S, Picard F, Rudolf G, Noe E, Achaz G, Thomas P, Genton P, Mundwiller E, Wolff M, Marescaux C, Miles R, Baulac M, Hirsch E, Leguern E, Baulac S (2013). Mutations of DEPDC5 cause autosomal dominant focal epilepsies. *Nat Genet* 45:552-555.
Jacquemont C, Taniguchi T (2007). The Fanconi anemia pathway and ubiquitin. *BMC Biochem* 8 Suppl 1:S10.
Jara-Prado A, Ochoa A, Alonso ME, Villeda GAL, Fernández-Valverde F, Ruano-Calderón L, Vargas-Cañas S, Durón RM, Delgado-Escueta AV, Martínez-Juárez IE (2014). Late onset Lafora disease and novel EPM2A mutations: Breaking paradigms. *Epilepsy Res* 108:1501-1510.
Joensuu T, Tegelberg S, Reinmaa E, Segerstråle M, Hakala P, Pehkonen H, Korpi ER, Tyynelä J, Taira T, Hovatta I, Kopra O, Lehesjoki AE (2014). Gene expression alterations in the cerebellum and granule neurons of Cstb(-/-) mouse are associated with early synaptic changes and inflammation. *PLoS ONE* 9:e89321.
Joho RH, Ho CS, Marks GA (1999). Increased gamma- and decreased delta-oscillations in a mouse deficient for a potassium channel expressed in fast-spiking interneurons. *J Neurophysiol* 82:1855-1864.
Kalachikov S, Evgrafov O, Ross B, Winawer M, Barker-Cummings C, Boneschi FM, Choi C, Morozov P, Das K, Teplitskaya E, Yu A, Cayanis E, Penchaszadeh G, Kottmann AH, Pedley TA, Hauser WA, Ottman R, Gilliam TC (2002). Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features. *Nat Genet* 30:335-341.
Kearney JA, Wiste AK, Stephani U, Trudeau MM, Siegel A, RamachandranNair R, Elterman RD, Muhle H, Reinsdorf J, Shields WD, Meisler MH, Escayg A (2006). Recurrent De Novo Mutations of SCN1A in Severe Myoclonic Epilepsy of Infancy. *Pediatr Neurol* 34:116-120.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler, David (2002). The Human Genome Browser at UCSC. *Genome Res* 12:996-1006.
Khan S, Al Baradie R (2012). Epileptic Encephalopathies: An Overview. *Epilepsy Res Treat* 2012:8.
Kimura M (1968). Evolutionary Rate at the Molecular Level. *Nature* 217:624-626.
Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46:310-315.
Kishino T, Lalande M, Wagstaff J (1997). UBE3A/E6-AP mutations cause Angelman syndrome. *Nat Genet* 15:70-73.
Knupp K, Wirrell E (2014). Progressive myoclonic epilepsies: It takes a village to make a diagnosis. *Neurology* 82:378-379.
Kollmann K, Uusi-Rauva K, Scifo E, Tyynelä J, Jalanko A, Braulke T (2013). Cell biology and function of neuronal ceroid lipofuscinosis-related proteins. *BBA-Mol Basis Dis* 1832:1866-1881.

Komatsu M, Chiba T, Tatsumi K, Iemura S, Tanida I, Okazaki N, Ueno T, Kominami E, Natsume T, Tanaka K (2004). A novel protein-conjugating system for Ufm1, a ubiquitin-fold modifier. *EMBO J* 23:1977-86.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488:471-475.

Krueger DA, Wilfong AA, Holland-Bouley K, Anderson AE, Agricola K, Tudor C, Mays M, Lopez CM, Kim MO, Franz DN (2013). Everolimus treatment of refractory epilepsy in tuberous sclerosis complex. *Ann Neurol* 74:679-87.

Kumar P, Henikoff S, Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073-1081.

Kälviäinen R (2015). Progressive Myoclonus Epilepsies. *Semin Neurol* 35:293-299.

Lai M, Huijbers MGM, Lancaster E, Graus F, Bataller L, Balice-Gordon R, Cowell JK, Dalmau J (2010). Investigation of LGI1 as the antigen in limbic encephalitis previously attributed to potassium channels: a case series. *Lancet Neurol* 9:776-785.

Lal D, Ruppert A-K, Trucks H, Schulz H, de Kovel CG, Kasteleijn-Nolst Trenité D, Sonsma ACM, Koeleman BP, Lindhout D, Weber YG, Lerche H, Kapser C, Schankin CJ, Kunz WS, Surges R, Elger CE, Gaus V, Schmitz B, Helbig I, Muhle H, Stephani U, Klein KM, Rosenow F, Neubauer BA, Reinthaler EM, Zimprich F, Feucht M, Møller RS, Hjalgrim H, De Jonghe P, Suls A, Lieb W, Franke A, Strauch K, Gieger C, Schurmann C, Schminke U, Nürnberg P, Sander T, Consortium E (2015). Burden Analysis of Rare Microdeletions Suggests a Strong Impact of Neurodevelopmental Genes in Genetic Generalised Epilepsies. *PLoS Genet* 11:e1005226.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 44:D862-D868.

Langmead B, Trapnell C, Pop M, Salzberg S (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

Lau D, de Miera EV-S, Contreras D, Ozaita A, Harvey M, Chow A, Noebels JL, Paylor R, Morgan JI, Leonard CS, Rudy B (2000). Impaired Fast-Spiking, Suppressed Cortical Inhibition, and Increased Susceptibility to Seizures in Mice Lacking Kv3.2 K+ Channel Proteins. *J Neurosci* 20:9071-9085.

Lehtinen MK, Tegelberg S, Schipper H, Su H, Zukor H, Manninen O, Kopra O, Joensuu T, Hakala P, Bonni A, Lehesjoki A-E (2009). Cystatin B Deficiency Sensitizes Neurons to Oxidative Stress in Progressive Myoclonus Epilepsy, EPM1. *J Neurosci* 29:5910-5915.

Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, O'Donnell-Luria A, Ware J, Hill A, Cummings B, Tukiainen T, Birnbaum D, Kosmicki J, Duncan L, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Cooper D, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki M, Levy Moonshine A, Natarajan P, Orozco L, Peloso G, Poplin R, Rivas M, Ruano-Rubio V, Ruderfer D, Shakir K, Stenson P, Stevens C, Thomas B, Tiao G, Tusie-Luna M, Weisburd B, Won H-H, Yu D, Altshuler D, Ardissino D, Boehnke M, Danesh J, Roberto E, Florez J, Gabriel S, Getz G, Hultman C, Kathiresan S, Laakso M, McCarroll S, McCarthy M, McGovern D, McPherson R, Neale B, Palotie A, Purcell S, Saleheen D, Scharf J, Sklar P, Patrick S, Tuomilehto J, Watkins H, Wilson J, Daly M, MacArthur D (2015). Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv doi:101101/030338*

Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein Coding Regions. *Hum Mutat* 36:815-822.

Lemaire K, Moura RF, Granvik M, Igoillo-Esteve M, Hohmeier HE, Hendrickx N, Newgard CB, Waelkens E, Cnop M, Schuit F (2011). Ubiquitin fold modifier 1 (UFM1) and its

target UFBP1 protect pancreatic beta cells from ER stress-induced apoptosis. *PLoS ONE* 6:e18517.

Lemke JR, Lal D, Reinthaler EM, Steiner I, Nothnagel M, Alber M, Geider K, Laube B, Schwake M, Finsterwalder K, Franke A, Schilhabel M, Jahn JA, Muhle H, Boor R, Van Paesschen W, Caraballo R, Fejerman N, Weckhuysen S, De Jonghe P, Larsen J, Moller RS, Hjalgrim H, Addis L, Tang S, Hughes E, Pal DK, Veri K, Vaher U, Talvik T, Dimova P, Lopez RG, Serratosa JM, Linnankivi T, Lehesjoki A-E, Ruf S, Wolff M, Buerki S, Wohlrab G, Kroell J, Datta AN, Fiedler B, Kurlemann G, Kluger G, Hahn A, Haberlandt DE, Kutzer C, Sperner J, Becker F, Weber YG, Feucht M, Steinbock H, Neophythou B, Ronen GM, Gruber-Sedlmayr U, Geldner J, Harvey RJ, Hoffmann P, Herms S, Altmuller J, Toliat MR, Thiele H, Nurnberg P, Wilhelm C, Stephani U, Helbig I, Lerche H, Zimprich F, Neubauer BA, Biskup S, von Spiczak S (2013). Mutations in GRIN2A cause idiopathic focal epilepsy with rolandic spikes. *Nat Genet* 45:1067-1072.

Lemke JR, Riesch E, Scheurenbrand T, Schubach M, Wilhelm C, Steiner I, Hansen J, Courage C, Gallati S, Bürki S, Strozzi S, Simonetti BG, Grunt S, Steinlin M, Alber M, Wolff M, Klopstock T, Prott EC, Lorenz R, Spaich C, Rona S, Lakshminarasimhan M, Kröll J, Dorn T, Krämer G, Synofzik M, Becker F, Weber YG, Lerche H, Böhm D, Biskup S (2012). Targeted next generation sequencing as a diagnostic tool in epileptic disorders. *Epilepsia* 53:1387-1398.

Lesca G, Rudolf G, Bruneau N, Lozovaya N, Labalme A, Boutry-Kryza N, Salmi M, Tsintsadze T, Addis L, Motte J, Wright S, Tsintsadze V, Michel A, Doummar D, Lascelles K, Strug L, Waters P, de Bellescize J, Vrielynck P, de Saint Martin A, Ville D, Ryvlin P, Arzimanoglou A, Hirsch E, Vincent A, Pal D, Burnashev N, Sanlaville D, Szepetowski P (2013). GRIN2A mutations in acquired epileptic aphasia and related childhood focal epilepsies and encephalopathies with speech and language dysfunction. *Nat Genet* 45:1061-1066.

Letts VA, Felix R, Biddlecome GH, Arikkath J, Mahaffey CL, Valenzuela A, Bartlett FS, Mori Y, Campbell KP, Frankel WN (1998). The mouse stargazer gene encodes a neuronal Ca2+-channel [gamma] subunit. *Nat Genet* 19:340-347.

Leutenegger A-L, Prum B, Génin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA (2003). Estimation of the Inbreeding Coefficient through Use of Genomic Data. *Am J Hum Genet* 73:516-523.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.

Lim JS, Kim W-i, Kang H-C, Kim SH, Park AH, Park EK, Cho Y-W, Kim S, Kim HM, Kim JA, Kim J, Rhee H, Kang S-G, Kim HD, Kim D, Kim D-S, Lee JH (2015). Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. *Nat Med* 21:395-400.

Liu H, Shim AHR, He X (2009). Structural characterization of the ectodomain of a disintegrin and metalloproteinase-22 (ADAM22), a neural adhesion receptor instead of metalloproteinase: insights on ADAM function. *J Biol Chem* 284:29077-29086.

Liu X, Han S, Wang Z, Gelernter J, Yang B-Z (2013). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE* 8:e75619.

Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Lee S, Chittenden TW, D'Gama AM, Cai X, Luquette LJ, Lee E, Park PJ, Walsh CA (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 350:94-98.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L,

Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struewing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45:580-585.

Loscher W, Klitgaard H, Twyman RE, Schmidt D (2013). New avenues for anti-epileptic drug discovery and development. *Nat Rev Drug Discov* 12:757-776.

Lovero KL, Fukata Y, Granger AJ, Fukata M, Nicoll RA (2015). The LGI1-ADAM22 protein complex directs synapse maturation through regulation of PSD-95 function. *Proc Natl Acad Sci U S A* 112:E4129-E4137.

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, Conrad DF, Cooper GM, Cox NJ, Daly MJ, Gerstein MB, Goldstein DB, Hirschhorn JN, Leal SM, Pennacchio LA, Stamatoyannopoulos JA, Sunyaev SR, Valle D, Voight BF, Winckler W, Gunter C (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature* 508:469-476.

Magaudda A, Ferlazzo E, Nguyen V-H, Genton P (2006). Unverricht-Lundborg Disease, a Condition with Self-limited Progression: Long-term Follow-up of 20 Patients. *Epilepsia* 47:860-866.

Mardis ER (2013). Next-Generation Sequencing Platforms. *Annu Rev Anal Chem* 6:287-303.

Marini C, Scheffer IE, Crossland KM, Grinton BE, Phillips FL, McMahon JM, Turner SJ, Dean JT, Kivity S, Mazarib A, Neufeld MY, Korczyn AD, Harkin LA, Dibbens LM, Wallace RH, Mulley JC, Berkovic SF (2004). Genetic architecture of idiopathic generalized epilepsy: clinical genetic analysis of 55 multiplex families. *Epilepsia* 45:467-78.

Markram H, Toledo-Rodriguez M, Wang Y, Gupta A, Silberberg G, Wu C (2004). Interneurons of the neocortical inhibitory system. *Nat Rev Neurosci* 5:793-807.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297-1303.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26:2069-2070.

McTague A, Howell KB, Cross JH, Kurian MA, Scheffer IE (2016). The genetic landscape of the epileptic encephalopathies of infancy and childhood. *Lancet Neurol* 15:304-316.

Mefford HC, Yendle SC, Hsu C, Cook J, Geraghty E, McMahon JM, Eeg-Olofsson O, Sadleir LG, Gill D, Ben-Zeev B, Lerman-Sagie T, Mackay M, Freeman JL, Andermann E, Pelakanos JT, Andrews I, Wallace G, Eichler EE, Berkovic SF, Scheffer IE (2011). Rare copy number variants are an important cause of epileptic encephalopathies. *Ann Neurol* 70:974-85.

Meisler MH OBJ (2012). Gene Interactions and Modifiers in Epilepsy. Jasper's Basic Mechanisms of the Epilepsies [Internet]. A. M. Noebels JL, Rogawski MA, et al. Bethesda (MD), National Center for Biotechnology Information (US).

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* 15:1-11.

Mikati MA, Jiang Y-h, Carboni M, Shashi V, Petrovski S, Spillmann R, Milligan CJ, Li M, Grefe A, McConkie A, Berkovic S, Scheffer I, Mullen S, Bonner M, Petrou S, Goldstein

D (2015). Quinidine in the treatment of KCNT1 positive epilepsies. *Ann Neurol* 78:995-999.
Milh M, Falace A, Villeneuve N, Vanni N, Cacciagli P, Assereto S, Nabbout R, Benfenati F, Zara F, Chabrol B, Villard L, Fassio A (2013). Novel Compound Heterozygous Mutations in TBC1D24 Cause Familial Malignant Migrating Partial Seizures of Infancy. *Hum Mutat* 34:869-872.
Milligan CJ, Li M, Gazina EV, Heron SE, Nair U, Trager C, Reid CA, Venkat A, Younkin DP, Dlugos DJ, Petrovski S, Goldstein DB, Dibbens LM, Scheffer IE, Berkovic SF, Petrou S (2014). KCNT1 gain of function in 2 epilepsy phenotypes is reversed by quinidine. *Ann Neurol* 75:581-590.
Minassian NA, Lin M-CA, Papazian DM (2012). Altered Kv3.3 channel gating in early-onset spinocerebellar ataxia type 13. *J Physiol* 590:1599-1614.
Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, McLean CY, Tung JY, Yu LPC, Gambetti P, Blevins J, Zhang S, Cohen Y, Chen W, Yamada M, Hamaguchi T, Sanjo N, Mizusawa H, Nakamura Y, Kitamoto T, Collins SJ, Boyd A, Will RG, Knight R, Ponto C, Zerr I, Kraus TFJ, Eigenbrod S, Giese A, Calero M, de Pedro-Cuesta J, Haïk S, Laplanche J-L, Bouaziz-Amar E, Brandel J-P, Capellari S, Parchi P, Poleggi A, Ladogana A, O'Donnell-Luria AH, Karczewski KJ, Marshall JL, Boehnke M, Laakso M, Mohlke KL, Kähler A, Chambert K, McCarroll S, Sullivan PF, Hultman CM, Purcell SM, Sklar P, van der Lee SJ, Rozemuller A, Jansen C, Hofman A, Kraaij R, van Rooij JGJ, Ikram MA, Uitterlinden AG, van Duijn CM, Daly MJ, MacArthur DG (2016). Quantifying prion disease penetrance using large population control cohorts. *Sci Transl Med* 8:322ra9-322ra9.
Myers CT, Mefford HC (2015). Advancing epilepsy genetics in the genomic era. *Genome Med* 7:1-11.
Németh AH, Kwasniewska AC, Lise S, Parolin Schnekenberg R, Becker EBE, Bera KD, Shanks ME, Gregory L, Buck D, Zameel Cader M, Talbot K, de Silva R, Fletcher N, Hastings R, Jayawant S, Morrison PJ, Worth P, Taylor M, Tolmie J, O'Regan M, Consortium UA, Valentine R, Packham E, Evans J, Seller A, Ragoussis J (2013). Next generation sequencing for molecular diagnosis of neurological disorders using ataxias as a model. *Brain* 136:3106-3118.
Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42:30-35.
Noachtar S, Rémi J (2009). The role of EEG in epilepsy: A critical review. *Epilepsy Behav* 15:22-33.
Noebels J (2015). Pathway-driven discovery of epilepsy genes. *Nat Neurosci* 18:344-350.
O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, Lyon G (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 5:28.
O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, Eichler EE (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485:246-250.
Ogawa Y, Oses-Prieto J, Kim MY, Horresh I, Peles E, Burlingame AL, Trimmer JS, Meijer D, Rasband MN (2010). ADAM22, a Kv1 channel-interacting protein, recruits membrane-associated guanylate kinases to juxtaparanodes of myelinated axons. *J Neurosci* 30:1038-1048.
Ohkawa T, Fukata Y, Yamasaki M, Miyazaki T, Yokoi N, Takashima H, Watanabe M, Watanabe O, Fukata M (2013). Autoantibodies to Epilepsy-Related LGI1 in Limbic Encephalitis Neutralize LGI1-ADAM22 Interaction and Reduce Synaptic AMPA Receptors. *J Neurosci* 33:18161-18174.

Okuneva O, Korber I, Li Z, Tian L, Joensuu T, Kopra O, Lehesjoki AE (2015). Abnormal microglial activation in the Cstb(-/-) mouse, a model for progressive myoclonus epilepsy, EPM1. *Glia* 63:400-11.

Orhan G, Bock M, Schepers D, Ilina EI, Reichel SN, Loffler H, Jezutkovic N, Weckhuysen S, Mandelstam S, Suls A, Danker T, Guenther E, Scheffer IE, De Jonghe P, Lerche H, Maljevic S (2014). Dominant-negative effects of KCNQ2 mutations are associated with epileptic encephalopathy. *Ann Neurol* 75:382-94.

Ortolano S, Vieitez I, Agis-Balboa RC, Spuch C (2014). Loss of GABAergic cortical neurons underlies the neuropathology of Lafora disease. *Mol Brain* 7:7.

Ottman R, Lipton RB (1994). Comorbidity of migraine and epilepsy. *Neurology* 44:2105-10.

Owuor K, Harel NY, Englot DJ, Hisama F, Blumenfeld H, Strittmatter SM (2009). LGI1-associated epilepsy through altered ADAM23-dependent neuronal morphology. *Mol Cell Neurosci* 42:448-457.

Ozaita A, Martone ME, Ellisman MH, Rudy B (2002). Differential subcellular localization of the two alternatively spliced isoforms of the Kv3.1 potassium channel subunit in brain. *J Neurophysiol* 88:394-408.

Ozcan L, Tabas I (2012). Role of Endoplasmic Reticulum Stress in Metabolic Disease and Other Disorders. *Annu Rev Med* 63:317-328.

Panayiotopoulos C (2005). Idiopathic Generalised Epilepsies. The Epilepsies: Seizures, Syndromes and Management. Oxfordshire (UK), Bladon Medical Publishing.

Peltonen L, Jalanko A, Varilo T (1999). Molecular genetics of the Finnish disease heritage. *Hum Mol Genet* 8:1913-1923.

Peltonen L, Perola M, Naukkarinen J, Palotie A (2006). Lessons from studying monogenic disease for common disease. *Hum Mol Genet* 15 Spec No 1:R67-74.

Pennacchio LA, Bouley DM, Higgins KM, Scott MP, Noebels JL, Myers RM (1998). Progressive ataxia, myoclonic epilepsy and cerebellar apoptosis in cystatin B-deficient mice. *Nat Genet* 20:251-8.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* 9:e1003709.

Pfeifer GP (2006). Mutagenesis at methylated CpG sequences. *Curr Top Microbiol Immunol* 301:259-81.

Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, Brunner HG, Buske OJ, Carey K, Doll C, Dumitriu S, Dyke SOM, den Dunnen JT, Firth HV, Gibbs RA, Girdea M, Gonzalez M, Haendel MA, Hamosh A, Holm IA, Huang L, Hurles ME, Hutton B, Krier JB, Misyura A, Mungall CJ, Paschall J, Paten B, Robinson PN, Schiettecatte F, Sobreira NL, Swaminathan GJ, Taschner PE, Terry SF, Washington NL, Züchner S, Boycott KM, Rehm HL (2015). The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Hum Mutat* 36:915-921.

Pierson TM, Yuan H, Marsh ED, Fuentes-Fajardo K, Adams DR, Markello T, Golas G, Simeonov DR, Holloman C, Tankovic A, Karamchandani MM, Schreiber JM, Mullikin JC, Tifft CJ, Toro C, Boerkoel CF, Traynelis SF, Gahl WA (2014). GRIN2A mutation and early-onset epileptic encephalopathy: personalized therapy with memantine. *Ann Clin Transl Neurol* 1:190-198.

Poduri A, Evrony GD, Cai X, Walsh CA (2013). Somatic Mutation, Genomic Variation, and Neurological Disease. *Science* 341.

Poultney Christopher S, Goldberg Arthur P, Drapeau E, Kou Y, Harony-Nicolas H, Kajiwara Y, De Rubeis S, Durand S, Stevens C, Rehnström K, Palotie A, Daly Mark J, Ma'ayan A, Fromer M, Buxbaum Joseph D (2013). Identification of Small Exonic CNV from Whole-Exome Sequence Data and Application to Autism Spectrum Disorder. *Am J Hum Genet* 93:607-619.

Powis Z, Farwell KD, Alamillo CL, Tang S (2016). Diagnostic exome sequencing for patients with a family history of consanguinity: over 38% of positive results are not autosomal recessive pattern. *J Hum Genet* 61:173-175.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559-575.

Purves D, Augustine GJ, Fitzpatrick D, Katz LC, LaMantia A-S, McNamara JO, Williams SM (2011). Summation of Synaptic Potentials. Neuroscience. Sunderland (MA), Sinauer Associates.

Rabbani B, Tekin M, Mahdieh N (2014). The promise of whole-exome sequencing in medical genetics. *J Hum Genet* 59:5-15.

Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, Dominiczak A, Morris A, Porteous D, Smith B, Stratton MR, Consortium UK, Hurles ME (2016). Timing, rates and spectra of human germline mutation. *Nat Genet* 48:126-133.

Ramser J, Ahearn ME, Lenski C, Yariz KO, Hellebrand H, von Rhein M, Clark RD, Schmutzler RK, Lichtner P, Hoffman EP, Meindl A, Baumbach-Reardon L (2008). Rare missense and synonymous variants in UBE1 are associated with X-linked infantile spinal muscular atrophy. *Am J Hum Genet* 82:188-93.

Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, Zetsche B, Shalem O, Wu X, Makarova KS, Koonin EV, Sharp PA, Zhang F (2015a). In vivo genome editing using Staphylococcus aureus Cas9. *Nature* 520:186-91.

Ran X, Li J, Shao Q, Chen H, Lin Z, Sun ZS, Wu J (2015b). EpilepsyGene: a genetic resource for genes and mutations related to epilepsy. *Nucleic Acids Res* 43:D893-9.

Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E (2013). ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 15:733-747.

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS (2015). ClinGen — The Clinical Genome Resource. *N Engl J Med* 372:2235-2242.

Rehman AU, Santos-Cortez RL, Morell RJ, Drummond MC, Ito T, Lee K, Khan AA, Basra MA, Wasif N, Ayub M, Ali RA, Raza SI, Nickerson DA, Shendure J, Bamshad M, Riazuddin S, Billington N, Khan SN, Friedman PL, Griffith AJ, Ahmad W, Leal SM, Friedman TB (2014). Mutations in TBC1D24, a gene associated with epilepsy, also cause nonsyndromic deafness DFNB86. *Am J Hum Genet* 94:144-152.

Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, Vertino-Bell A, Smaoui N, Neidich J, Monaghan KG, McKnight D, Bai R, Suchy S, Friedman B, Tahiliani J, Pineda-Alvarez D, Richard G, Brandt T, Haverfield E, Chung WK, Bale S (2015). Clinical application of whole-exome sequencing across clinical indications. *Genet Med*.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17:405-423.

Ried T, Rudy B, de Miera EV-S, Lau D, Ward DC, Sen K (1993). Localization of a Highly Conserved Human Potassium Channel Gene (NGK2-KV4; KCNC1) to Chromosome 11p15. *Genomics* 15:405-411.

Riggs AJ, Riggs JE (2005). Epilepsy's Role in the Historical Differentiation of Religion, Magic, and Science. *Epilepsia* 46:452-453.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011). Integrative genomics viewer. *Nat Biotechnol* 29:24-26.

Rosanoff MJ, Ottman R (2008). Penetrance of LGI1 mutations in autosomal dominant partial epilepsy with auditory features. *Neurology* 71:567-571.

Rosato-Siri MD, Zambello E, Mutinelli C, Garbati N, Benedetti R, Aldegheri L, Graziani F, Virginio C, Alvaro G, Large CH (2015). A Novel Modulator of Kv3 Potassium Channels Regulates the Firing of Parvalbumin-Positive Cortical Interneurons. *J Pharmacol Exp Ther* 354:251-260.

Rousseau F, Bonaventure J, Legeai-Mallet L, Pelet A, Rozet J-M, Maroteaux P, Merrer ML, Munnich A (1994). Mutations in the gene encoding fibroblast growth factor receptor-3 in achondroplasia. *Nature* 371:252-254.

Rudy B, McBain CJ (2001). Kv3 channels: voltage-gated K+ channels designed for high-frequency repetitive firing. *Trends Neurosci* 24:517-526.

Sagane K, Hayakawa K, Kai J, Hirohashi T, Takahashi E, Miyamoto N, Ino M, Oki T, Yamazaki K, Nagasu T (2005). Ataxia and peripheral nerve hypomyelination in ADAM22-deficient mice. *BMC Neurosci* 6:33.

Saitsu H, Kato M, Mizuguchi T, Hamada K, Osaka H, Tohyama J, Uruno K, Kumada S, Nishiyama K, Nishimura A, Okada I, Yoshimura Y, Hirai S-i, Kumada T, Hayasaka K, Fukuda A, Ogata K, Matsumoto N (2008). De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. *Nat Genet* 40:782-788.

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, Wall DP, MacArthur DG, Gabriel SB, dePristo M, Purcell SM, Palotie A, Boerwinkle E, Buxbaum JD, Edwin H. Cook J, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Neale BM, Daly MJ (2014). A framework for the interpretation of *de novo* mutation in human disease. *Nat Genet* 46:944-950.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74:5463-7.

Scharfman HE (2007). The neurobiology of epilepsy. *Curr Neurol Neurosci Rep* 7:348-354.

Schwarz JM, Cooper DN, Schuelke M, Seelow D (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Meth* 11:361-362.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 7:575-576.

Sekirnjak C, Martone ME, Weiser M, Deerinck T, Bueno E, Rudy B, Ellisman M (1997). Subcellular localization of the K+ channel subunit Kv3.1b in selected rat CNS neurons. *Brain Res* 766:173-87.

Seoh S-A, Sigg D, Papazian DM, Bezanilla F (1996). Voltage-Sensing Residues in the S2 and S4 Segments of the Shaker K+ Channel. *Neuron* 16:1159-1167.

Shahwan A, Farrell M, Delanty N (2005). Progressive myoclonic epilepsies: a review of genetic and therapeutic aspects. *Lancet Neurol* 4:239-248.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005). Segmental Duplications and Copy-Number Variation in the Human Genome. *Am J Hum Genet* 77:78-88.

Shi L, Li B, Huang Y, Ling X, Liu T, Lyon GJ, Xu A, Wang K (2014). “Genotype-first” approaches on a curious case of idiopathic progressive cognitive decline. *BMC Medical Genomics* 7:1-10.

Shiang R, Thompson LM, Zhu Y-Z, Church DM, Fielder TJ, Bocian M, Winokur ST, Wasmuth JJ (1994). Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. *Cell* 78:335-342.

Shinde DN, Elmer DP, Calabrese P, Boulanger J, Arnheim N, Tiemann-Boege I (2013). New evidence for positive selection helps explain the paternal age effect observed in achondroplasia. *Hum Mol Genet* 22:4117-4126.

Shire. (2016). "Rare Disease Impact Report https://globalgenes.org/wp-content/uploads/2013/04/ShireReport-1.pdf, Accessed [January 20, 2016]."

Shoffner JM, Lott MT, Lezza AM, Seibel P, Ballinger SW, Wallace DC (1990). Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNA(Lys) mutation. *Cell* 61:931-937.

Simons C, Wolf Nicole I, McNeil N, Caldovic L, Devaney Joseph M, Takanohashi A, Crawford J, Ru K, Grimmond Sean M, Miller D, Tonduti D, Schmidt Johanna L, Chudnow Robert S, van Coster R, Lagae L, Kisler J, Sperner J, van der Knaap Marjo S, Schiffmann R, Taft Ryan J, Vanderver A (2013). A De Novo Mutation in the Beta-

Tubulin Gene TUBB4A Results in the Leukoencephalopathy Hypomyelination with Atrophy of the Basal Ganglia and Cerebellum. *Am J Hum Genet* 92:767-773.

Singh NA, Charlier C, Stauffer D, DuPont BR, Leach RJ, Melis R, Ronen GM, Bjerre I, Quattlebaum T, Murphy JV, McHarg ML, Gagnon D, Rosales TO, Peiffer A, Anderson VE, Leppert M (1998). A novel potassium channel gene, KCNQ2, is mutated in an inherited epilepsy of newborns. *Nat Genet* 18:25-29.

Sisodiya S. (2015). "Genetics of the epilepsies." from http://www.epilepsysociety.org.uk/sites/default/files/attachments/Chapter05Sisodiya2015.pdf.

Slaymaker IM, Gao L, Zetsche B, Scott DA, Yan WX, Zhang F (2016). Rationally engineered Cas9 nucleases with improved specificity. *Science* 351:84-8.

Soden SE, Saunders CJ, Willig LK, Farrow EG, Smith LD, Petrikin JE, LePichon J-B, Miller NA, Thiffault I, Dinwiddie DL, Twist G, Noll A, Heese BA, Zellmer L, Atherton AM, Abdelmoity AT, Safina N, Nyp SS, Zuccarelli B, Larson IA, Modrcin A, Herd S, Creed M, Ye Z, Yuan X, Brodsky RA, Kingsmore SF (2014). Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Sci Transl Med* 6:265ra168-265ra168.

Srivastava S, Cohen JS, Vernon H, Barañano K, McClellan R, Jamal L, Naidu S, Fatemi A (2014). Clinical whole exome sequencing in child neurology practice. *Ann Neurol* 76:473-483.

Staley K (2015). Molecular mechanisms of epilepsy. *Nat Neurosci* 18:367-372.

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S (2015). Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* 200:413-422.

Stark Z, Tan TY, Chong B, Brett GR, Yap P, Walsh M, Yeung A, Peters H, Mordaunt D, Cowie S, Amor DJ, Savarirayan R, McGillivray G, Downie L, Ekert PG, Theda C, James PA, Yaplito-Lee J, Ryan MM, Leventer RJ, Creed E, Macciocca I, Bell KM, Oshlack A, Sadedin S, Georgeson P, Anderson C, Thorne N, Melbourne Genomics Health A, Gaff C, White SM (2016). A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet Med*.

Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, Nalpathamkalam T, Pellecchia G, Yuen RKC, Szego MJ, Hayeems RZ, Shaul RZ, Brudno M, Girdea M, Frey B, Alipanahi B, Ahmed S, Babul-Hirji R, Porras RB, Carter MT, Chad L, Chaudhry A, Chitayat D, Doust SJ, Cytrynbaum C, Dupuis L, Ejaz R, Fishman L, Guerin A, Hashemi B, Helal M, Hewson S, Inbar-Feigenberg M, Kannu P, Karp N, Kim RH, Kronick J, Liston E, MacDonald H, Mercimek-Mahmutoglu S, Mendoza-Londono R, Nasr E, Nimmo G, Parkinson N, Quercia N, Raiman J, Roifman M, Schulze A, Shugar A, Shuman C, Sinajon P, Siriwardena K, Weksberg R, Yoon G, Carew C, Erickson R, Leach RA, Klein R, Ray PN, Meyn MS, Scherer SW, Cohn RD, Marshall CR (2016). Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *Npj Genomic Medicine* 1:15012.

Steinlein OK, Mulley JC, Propping P, Wallace RH, Phillips HA, Sutherland GR, Scheffer IE, Berkovic SF (1995). A missense mutation in the neuronal nicotinic acetylcholine receptor [alpha]4 subunit is associated with autosomal dominant nocturnal frontal lobe epilepsy. *Nat Genet* 11:201-203.

Stessman HA, Bernier R, Eichler EE (2014). A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell* 156:872-877.

Stranger BE, Stahl EA, Raj T (2011). Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics* 187:367-383.

Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, Hjartarson E, Sigurdsson GT, Jonasdottir A, Jonasdottir A, Sigurdsson A, Magnusson OT, Kong A, Helgason A, Holm H, Thorsteinsdottir U, Masson G, Gudbjartsson DF, Stefansson K (2015). Identification of a large set of rare complete human knockouts. *Nat Genet* 47:448-452.

Sulonen A-M, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C, Joensuu H, Raivio T, Suomalainen A, Saarela J

(2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* 12:R94.

Suls A, Jaehn Johanna A, Kecskés A, Weber Y, Weckhuysen S, Craiu Dana C, Siekierska A, Djémié T, Afrikanova T, Gormley P, von Spiczak S, Kluger G, Iliescu Catrinel M, Talvik T, Talvik I, Meral C, Caglayan Hande S, Giraldez Beatriz G, Serratosa J, Lemke Johannes R, Hoffman-Zacharska D, Szczepanik E, Barisic N, Komarek V, Hjalgrim H, Møller Rikke S, Linnankivi T, Dimova P, Striano P, Zara F, Marini C, Guerrini R, Depienne C, Baulac S, Kuhlenbäumer G, Crawford Alexander D, Lehesjoki A-E, de Witte Peter AM, Palotie A, Lerche H, Esguerra Camila V, De Jonghe P, Helbig I (2013). De Novo Loss-of-Function Mutations in CHD2 Cause a Fever-Sensitive Myoclonic Epileptic Encephalopathy Sharing Features with Dravet Syndrome. *Am J Hum Genet* 93:967-975.

Swiech L, Heidenreich M, Banerjee A, Habib N, Li Y, Trombetta J, Sur M, Zhang F (2015). In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nat Biotechnol* 33:102-6.

Synofzik M, Soehn A, Gburek-Augustat J, Schicks J, Karle K, Schule R, Haack T, Schoning M, Biskup S, Rudnik-Schoneborn S, Senderek J, Hoffmann K-T, MacLeod P, Schwarz J, Bender B, Kruger S, Kreuz F, Bauer P, Schols L (2013). Autosomal recessive spastic ataxia of Charlevoix Saguenay (ARSACS): expanding the genetic, clinical and imaging spectrum. *Orphanet J Rare Dis* 8:41.

Syrbe S, Hedrich UBS, Riesch E, Djemie T, Muller S, Moller RS, Maher B, Hernandez-Hernandez L, Synofzik M, Caglayan HS, Arslan M, Serratosa JM, Nothnagel M, May P, Krause R, Loffler H, Detert K, Dorn T, Vogt H, Kramer G, Schols L, Mullis PE, Linnankivi T, Lehesjoki A-E, Sterbova K, Craiu DC, Hoffman-Zacharska D, Korff CM, Weber YG, Steinlin M, Gallati S, Bertsche A, Bernhard MK, Merkenschlager A, Kiess W, Euro ER, Gonzalez M, Zuchner S, Palotie A, Suls A, De Jonghe P, Helbig I, Biskup S, Wolff M, Maljevic S, Schule R, Sisodiya SM, Weckhuysen S, Lerche H, Lemke JR (2015). De novo loss- or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. *Nat Genet* 47:393-399.

Tammaro P, Shimomura K, Proks P (2009). Xenopus Oocytes as a Heterologous Expression System for Studying Ion Channels with the Patch-Clamp Technique. Potassium Channels. J. Lippiat, Humana Press. **491:**127-139.

Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, Jiang Q, Allen AS, Zhu M (2014). An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Hum Mutat* 35:899-907.

Taskin B, von Schoubye NL, Sheykhzade M, Bastlund JF, Grunnet M, Jespersen T (2015). Biophysical characterization of KV3.1 potassium channel activating compounds. *Eur J Pharmacol* 758:164-170.

Tatsumi K, Sou YS, Tada N, Nakamura E, Iemura S, Natsume T, Kang SH, Chung CH, Kasahara M, Kominami E, Yamamoto M, Tanaka K, Komatsu M (2010). A novel type of E3 ligase for the Ufm1 conjugation system. *J Biol Chem* 285:5417-27.

Tatsumi K, Yamamoto-Mukai H, Shimizu R, Waguri S, Sou Y-S, Sakamoto A, Taya C, Shitara H, Hara T, Chung CH, Tanaka K, Yamamoto M, Komatsu M (2011). The Ufm1-activating enzyme Uba5 is indispensable for erythroid differentiation in mice. *Nat Commun* 2:181.

Taylor JC, Martin HC, Lise S, Broxholme J, Cazier J-B, Rimmer A, Kanapin A, Lunter G, Fiddy S, Allan C, Aricescu AR, Attar M, Babbs C, Becq J, Beeson D, Bento C, Bignell P, Blair E, Buckle VJ, Bull K, Cais O, Cario H, Chapel H, Copley RR, Cornall R, Craft J, Dahan K, Davenport EE, Dendrou C, Devuyst O, Fenwick AL, Flint J, Fugger L, Gilbert RD, Goriely A, Green A, Greger IH, Grocock R, Gruszczyk AV, Hastings R, Hatton E, Higgs D, Hill A, Holmes C, Howard M, Hughes L, Humburg P, Johnson D, Karpe F, Kingsbury Z, Kini U, Knight JC, Krohn J, Lamble S, Langman C, Lonie L, Luck J, McCarthy D, McGowan SJ, McMullin MF, Miller KA, Murray L, Nemeth AH, Nesbit MA, Nutt D, Ormondroyd E, Oturai AB, Pagnamenta A, Patel SY, Percy M, Petousi N, Piazza P, Piret SE, Polanco-Echeverry G, Popitsch N, Powrie F, Pugh C, Quek L, Robbins PA, Robson K, Russo A, Sahgal N, van Schouwenburg PA, Schuh A,

Silverman E, Simmons A, Sorensen PS, Sweeney E, Taylor J, Thakker RV, Tomlinson I, Trebes A, Twigg SRF, Uhlig HH, Vyas P, Vyse T, Wall SA, Watkins H, Whyte MP, Witty L, Wright B, Yau C, Buck D, Humphray S, Ratcliffe PJ, Bell JI, Wilkie AOM, Bentley D, Donnelly P, McVean G (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet* 47:717-726.

Tegelberg S, Kopra O, Joensuu T, Cooper JD, Lehesjoki AE (2012). Early microglial activation precedes neuronal loss in the brain of the Cstb-/- mouse model of progressive myoclonus epilepsy, EPM1. *J Neuropathol Exp Neurol* 71:40-53.

The 1000 Genomes Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073.

The 1000 Genomes Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56-65.

The Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 519:223-8.

The UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42:D191-D198.

Tian C, Zhu R, Zhu L, Qiu T, Cao Z, Kang T (2014). Potassium Channels: Structures, Diseases, and Modulators. *Chem Biol Drug* 83:1-26.

Tripathy D, Harnden K, Blackwell K, Robson M (2014). Next generation sequencing and tumor mutation profiling: are we ready for routine use in the oncology clinic? *BMC Med* 12:1-8.

Vadlamudi L, Dibbens LM, Lawrence KM, Iona X, McMahon JM, Murrell W, Mackay-Sim A, Scheffer IE, Berkovic SF (2010). Timing of de novo mutagenesis--a twin study of sodium-channel mutations. *N Engl J Med* 363:1335-40.

Wagnon JL, Meisler MH (2015). Recurrent and non-recurrent mutations of SCN8A in epileptic encephalopathy. *Front Neurol* 6.

Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL, McKenna A, Sivachenko A, Rosenberg M, Kiezun A, Voet D, Lawrence M, Lichtenstein LT, Gentry JG, Huang FW, Fostel J, Farlow D, Barbie D, Gandhi L, Lander ES, Gray SW, Joffe S, Janne P, Garber J, MacConaill L, Lindeman N, Rollins B, Kantoff P, Fisher SA, Gabriel S, Getz G, Garraway LA (2014). Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 20:682-688.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013). From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit Best Practices pipeline. *Curr Protoc Bioinform* 43:11.10.1-11.10.33.

Van Driest SL, Wells QS, Stallings S, et al. (2016). Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *JAMA* 315:47-57.

Wang G-S, Cooper TA (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 8:749-761.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17:1665-1674.

Wang K, Li M, Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164.

Waters MF, Minassian NA, Stevanin G, Figueroa KP, Bannister JPA, Nolte D, Mock AF, Evidente VGH, Fee DB, Muller U, Durr A, Brice A, Papazian DM, Pulst SM (2006). Mutations in voltage-gated potassium channel KCNC3 cause degenerative and developmental central nervous system phenotypes. *Nat Genet* 38:447-451.

Watson CM, Crinnion LA, Gleghorn L, Newman WG, Ramesar R, Beighton P, Wallis GA (2015). Identification of a mutation in the ubiquitin-fold modifier 1-specific peptidase 2 gene, UFSP2, in an extended South African family with Beukes hip dysplasia. *S Afr Med J* 105:558-63.

Weckhuysen S, Mandelstam S, Suls A, Audenaert D, Deconinck T, Claes LR, Deprez L, Smets K, Hristova D, Yordanova I, Jordanova A, Ceulemans B, Jansen A, Hasaerts D, Roelens F, Lagae L, Yendle S, Stanley T, Heron SE, Mulley JC, Berkovic SF, Scheffer IE, de Jonghe P (2012). KCNQ2 encephalopathy: emerging phenotype of a neonatal epileptic encephalopathy. *Ann Neurol* 71:15-25.

Veltman JA, Brunner HG (2012). De novo mutations in human genetic disease. *Nat Rev Genet* 13:565-575.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji R-R, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang ZY, Wang A, Wang X, Wang J, Wei M-H, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers Y-H, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang Y-H, Coyne M, Dahlke C, Mays AD, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001). The Sequence of the Human Genome. *Science* 291:1304-1351.

Wetterstrand K. (2016). "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed [20 Jan 2016].".

Wilkie AOM, Slaney SF, Oldridge M, Poole MD, Ashworth GJ, Hockley AD, Hayward RD, David DJ, Pulleyn LJ, Rutland P, Malcolm S, Winter RM, Reardon W (1995). Apert syndrome results from localized mutations of FGFR2 and is allelic with Crouzon syndrome. *Nat Genet* 9:165-172.

Willig LK, Petrikin JE, Smith LD, Saunders CJ, Thiffault I, Miller NA, Soden SE, Cakici JA, Herd SM, Twist G, Noll A, Creed M, Alba PM, Carpenter SL, Clements MA, Fischer RT, Hays JA, Kilbride H, McDonough RJ, Rosterman JL, Tsai SL, Zellmer L, Farrow EG, Kingsmore SF (2015). Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. *Lancet Respir Med* 3:377-387.

Vissers LELM, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BWM, Hoischen A, de Vries BBA, Brunner HG, Veltman JA (2010). A de novo paradigm for mental retardation. *Nat Genet* 42:1109-1112.

Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, Gogos JA, Karayiorgou M (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* 43:864-868.

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S (2012). FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE* 7:e52249.

Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C, Veeraraghavan N, Hawes A, Chiang T, Leduc M, Beuten J, Zhang J, He W, Scull J, Willis A, Landsverk M, Craigen WJ, Bekheirnia MR, Stray-Pedersen A, Liu P, Wen S, Alcaraz W, Cui H, Walkiewicz M, Reid J, Bainbridge M, Patel A, Boerwinkle E, Beaudet AL, Lupski JR, Plon SE, Gibbs RA, Eng CM (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 312:1870-9.

Yasuda T, Cuny H, Adams DJ (2013). Kv3.1 channels stimulate adult neural precursor cell proliferation and neuronal differentiation. *J Physiol* 591:2579-2591.

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134.

Yoo HM, Kang SH, Kim JY, Lee JE, Seong MW, Lee SW, Ka SH, Sou YS, Komatsu M, Tanaka K, Lee ST, Noh DY, Baek SH, Jeon YJ, Chung CH (2014). Modification of ASC1 by UFM1 is crucial for ERalpha transactivation and breast cancer development. *Mol Cell* 56:261-74.

Young K, Clark HB, Piccardo P, Dlouhy SR, Ghetti B (1997). Gerstmann–Sträussler–Scheinker disease with the PRNP P102L mutation and valine at codon 129. *Brain Res Mol Brain Res* 44:147-150.

Yu FH, Mantegazza M, Westenbroek RE, Robbins CA, Kalume F, Burton KA, Spain WJ, McKnight GS, Scheuer T, Catterall WA (2006). Reduced sodium current in GABAergic interneurons in a mouse model of severe myoclonic epilepsy in infancy. *Nat Neurosci* 9:1142-1149.

Yu YE, Wen L, Silva J, Li Z, Head K, Sossey-Alaoui K, Pao A, Mei L, Cowell JK (2010). Lgi1 null mutant mice exhibit myoclonic seizures and CA1 neuronal hyperexcitability. *Hum Mol Genet* 19:1702-1711.

Zhang Y, Zhang M, Wu J, Lei G, Li H (2012). Transcriptional regulation of the Ufm1 conjugation system in response to disturbance of the endoplasmic reticulum homeostasis and inhibition of vesicle trafficking. *PLoS ONE* 7:e48587.

Zhao J, Zhu J, Thornhill WB (2013). Spinocerebellar ataxia-13 Kv3.3 potassium channels: arginine-to-histidine mutations affect both functional and protein expression on the cell surface. *Biochem J* 454:259-265.

Zhao L, Longo-Guess C, Harris BS, Lee J-W, Ackerman SL (2005). Protein accumulation and neurodegeneration in the woozy mutant mouse is caused by disruption of SIL1, a cochaperone of BiP. *Nat Genet* 37:974-979.

Zhou Y-D, Lee S, Jin Z, Wright M, Smith SEP, Anderson MP (2009). Arrested maturation of excitatory synapses in autosomal dominant lateral temporal lobe epilepsy. *Nat Med* 15:1208-1214.

Özkaynak E, Abello G, Jaegle M, van Berge L, Hamer D, Kegel L, Driegen S, Sagane K, Bermingham JR, Meijer D (2010). Adam22 Is a Major Neuronal Receptor for Lgi4-Mediated Schwann Cell Signaling. *J Neurosci* 30:3857-3864.

10 ORIGINAL PUBLICATIONS

Recent Publications in this Series

24/2016 Niina Markkula
Prevalence, Predictors and Prognosis of Depressive Disorders in the General Population
25/2016 Katri Kallio
The Roles of Template RNA and Replication Proteins in the Formation of Semliki Forest Virus Replication Spherules
26/2015 Hanna Paatela
Role of Dehydroepiandrosterone in High-Density Lipoprotein-Mediated Vasodilation and in Adipose Tissue Steroid Biosynthesis
27/2016 Johanna Mäkelä
Neuroprotective Effects of PGC-1α Activators in Dopaminergic Neurons
28/2016 Sandra Söderholm
Phosphoproteomic Characterization of Viral Infection
29/2016 Mariann Lassenius
Bacterial Endotoxins in Type 1 Diabetes
30/2016 Mette Ilander
T and NK Cell Mediated Immunity in Chronic Myeloid Leukaemia
31/2016 Ninja Karikoski
The Prevalence and Histopathology of Endocrinopathic Laminitis in Horses
32/2016 Michael Backlund
Regulation of Angiotensin II Type 1 Receptor by Its Messenger RNA-Binding Proteins
33/2016 Stanislav Rozov
Circadian and Histaminergic Regulation of the Sleep-Wakefulness Cycle
34/2016 Bárbara Herranz Blanco
Multi-Approach Design and Fabrication of Hybrid Composites for Drug Delivery and Cancer Therapy
35/2016 Siri Tähtinen
Combining Oncolytic Immunotherapies to Break Tumor Resistance
36/2016 Katri Sääksjärvi
Diet, Lifestyle Factors, Metabolic Health and Risk of Parkinson's Disease – A Prospective Cohort Study
37/2016 Alexandre Borrel
Development of Computational Methods to Predict Protein Pocket Druggability and Profile Ligands using Structural Data
38/2016 Gonçalo Barreto
Innate Immunity in Osteoarthritis: The Role of Toll-Like Receptors and Cartilage Derived Mediators in the Disease Progression
39/2016 Alexey Yukin
Animal Models of Early Brain Disorders: Behavioural and Electrophysiological Approaches
40/2016 Satu Valo
Western Diet and Genetic Predisposition as Risk Factors of Colon Cancer
41/2016 Riccardo Siligato
Hormonal Regulation of Primary and Secondary Growth in the Root of *Arabidopsis thaliana*
42/2016 Janne Tynell
Virus-Host Interactions of Emerging Respiratory Pathogens
43/2016 Katri Korpela
Intestinal Microbiota Development in Childhood: Implications for Health and Disease

Helsinki 2016 ISSN 2342-3161 ISBN 978-951-51-2255-1